Univerzita Karlova v Praze
Přírodovědecká fakulta

Studijní program: Speciální chemicko-biologické obory
Studijní obor: Molekulární biologie a biochemie organismů

Albert Sokol

Ab initio predikce struktury membránových proteinů

Ab initio prediction of the membrane protein structures

Bakalářská práce

Školitel: RNDr. Radovan Fišer, Ph.D

Praha, 2016

Prohlášení:

Prohlašuji, že jsem závěrečnou práci zpracoval samostatně a že jsem uvedl všechny použité informační zdroje a literaturu. Tato práce ani její podstatná část nebyla předložena k získání jiného nebo stejného akademického titulu.

V Praze, 1. 5. 2016

Albert Sokol

Děkuji svému školiteli RNDr. Radovanu Fišerovi, Ph.D. za odborné konzultace a podporu.

Obsah

1 Abstrakt

Znalost trojrozměrné struktury proteinu je extrémně důležitá pro plné pochopení jeho funkce a molekulárních interakcí. Struktura je typicky určována experimentálně pomocí X-ray krystalografie a NMR spektroskopie, bohužel membránové proteiny často znamenají pro tyto metody vážný problém. Východiskem je výpočetní predikce na základě již zjištěných dat.

Ab initio predikce trojrozměrných modelů membránových proteinů je komplexní proces, který nevyužívá žádnou dostupnou strukturu proteinu jako celkovou šablonu. Existuje pár softwarů, které se tímto procesem zabývají a vybrané čtyři jsou detailně popsány v této práci. Jedná se o dva programy pro predikci transmembránových helikálních proteinů (Rosetta, EVfold_membrane) a dva pro predikci transmembránových beta barelů (EVfold_bb, 3D-SPOT).

Hlavní přístupy, které jsou využívány v predikci trojrozměrné struktury proteinu, jsou vkládání krátkých úseků sekvence aminokyselin, které jsou odvozené ze zjištěných proteinových struktur (Rosetta), využívání evoluční informace z mnoha jiných sekvencí proteinů (EVfold) a tvorba beta barelové domény na základě kombinování sousedních antiparalelních beta řetězců (3D-SPOT). Každý software používá různé externí programy na řešení specifických problémů, jako například predikce transmembránových úseků z pouhé sekvence nebo programy na dohledávání homologních sekvencí.

I přes množství problémů udělaly predikční programy značný vývoj a s rostoucím výpočetním výkonem a lepšími algoritmy bude jistě zajímavé sledovat, kam bude v budoucnu predikce směřovat. Cílem práce je detailní náhled na používané metody a přiblížit čtenářům někdy těžko pochopitelné algoritmy.

Klíčová slova

Predikce struktury membránových proteinů, *ab initio* predikce, Rosetta, EVfold, 3D-SPOT

Seznam zkratek

CO	Contact order	Řád interakcí
RMSD	Root-mean-square distance	Střední kvadratická vzdálenost
TM	Transmembrane	Transmembránový

2 Abstract

Knowledge of the three dimensional structure of the protein is extremely important for a full understanding of its function and molecular proteins interaction. The structure is typically determined experimentally by X-ray crystallography and NMR spectroscopy, unfortunately membrane proteins provide numerous problems for these methods. A possible solution is the computational prediction.

Ab initio prediction of three-dimensional models of the membrane proteins is a complex process which cannot use any available protein structure as a general template. There are few softwares that deal with this process and selected four are described in detail in this work. These are two programs for the prediction of transmembrane helical proteins (Rosetta, EVfold_membrane) and two for the prediction of transmembrane beta barrels (EVfold_bb, 3D-SPOT).

The main approaches that are used in the prediction of the three-dimensional structure of a protein are inserting short segments of amino acid sequences which are derived from the determined protein structures (Rosetta), using evolutionary information from many other protein sequences (EVfold) and formation of the beta barrel domains based on combining adjacent antiparallel beta chains (3D-SPOT). Every software uses a variety of external programs to address specific problems, such as the prediction of transmembrane segments of sequence or programs for finding homologous sequences.

Despite a number of problems the prediction programs have made considerable progress and with increasing computing power and better algorithms it will be certainly interesting to see where the future prediction will lead. The goal of this work is a detailed look at the used methods to introduce readers to sometimes complicated algorithms.

Keywords

Structure prediction of membrane proteins, *ab initio* prediction, Rosetta, EVfold, 3D-SPOT

3 Úvod

Programy predikující trojrozměrnou strukturu proteinu vznikaly jako doplněk experimentálních metod, jako jsou X-ray krystalografie a NMR spektroskopie. Za posledních 15 let udělaly predikční metody značný vývoj a většinu malých solubilních proteinů dokáží určit s velkou přesností. Oproti tomu se membránovými proteiny zabývá jen malá část z nich, a to hlavně z důvodu složitější predikce v anizotropickém membránovém prostředí než v roztoku. Další problém je, že membránové proteiny jsou často velké a se složitější topologií.

Predikce se obecně dělí na *ab initio* (od začátku) a homologní modelování, které využívá k tvorbě 3D modelu jiný protein s vyřešenou strukturou jako šablonu. Pokud žádná šablona není k nalezení, je použita *ab initio* predikce, která je sice náročnější, ale nezávislá na experimentálně zjištěných proteinech.

Samotná znalost trojrozměrné struktury proteinu je esenciální pro plné pochopení jeho molekulární funkce a meziproteinové interakce. Identifikování aktivních míst membránových proteinů může také urychlit vývoj nových léků.

Moje práce vychází z požadavků skupiny Fyziologie bakterií na programátorskou úpravu predikčního softwaru Rosetta, která je nyní často využívána pro predikci bakteriálních membránových proteinů. Práce je tedy zaměřena na detailní náhled do algoritmů a přístupů, které software Rosetta používá. Navíc Rosetta je první predikční nástroj, který se začal zabývat membránovými proteiny, a do členské základny, která ji vyvíjí, patří mnoho vědeckých skupin. Dále má volně přístupný zdrojový kód, který je možný upravit podle aktuálních potřeb. Mimo Rosettu je v této práci popsán program EVfold_membrane, který využívá evoluční informaci korelujících mutací k tvorbě 3D modelů helikálních membránových proteinů. V poslední části popisuji dva programy pro predikci membránových beta barelů, EVfold_bb a 3D-SPOT.

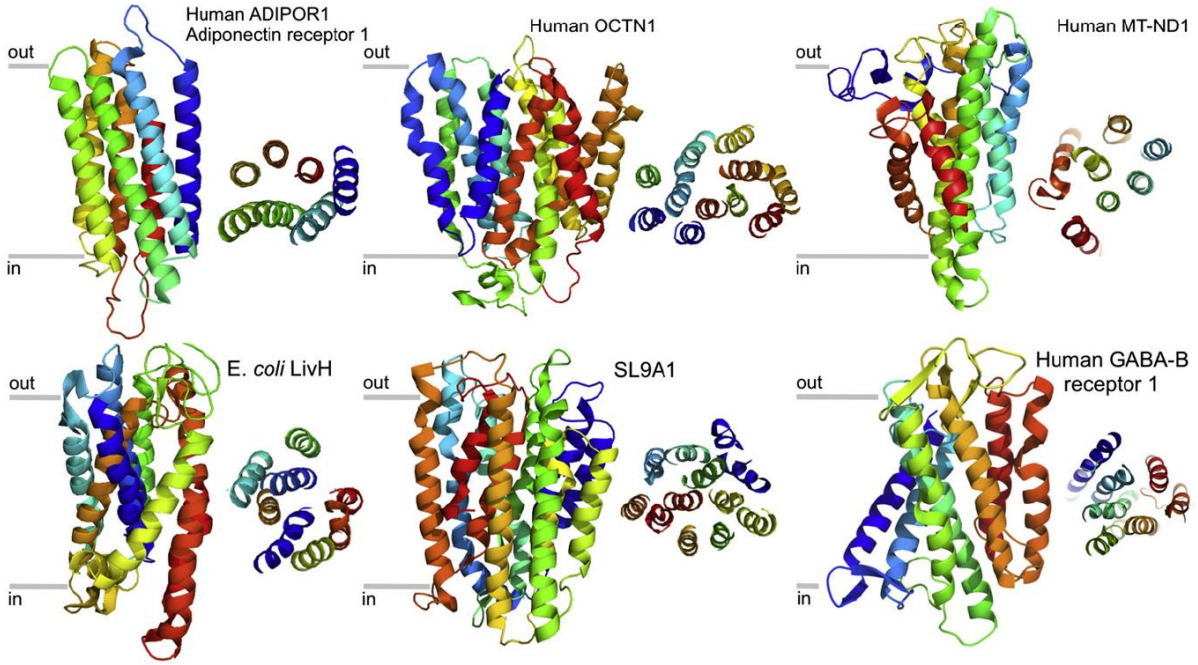
Cílem této práce je:

Do hloubky popsat reprezentativní programy pro *ab initio* predikci trojrozměrné struktury transmembránových proteinů a jeden si osvojit pro další výzkum, tedy úpravy algoritmů a dodatečnou analýzu predikovaných modelů, jako je například klastrování.

4 Membránové proteiny

Jako membránové proteiny jsou v predikčních metodách chápany především transmembránové (TM) proteiny, které procházejí skrz celou membránu a jsou její permanentní součástí. Predikce se soustřeďuje na dva hlavní strukturní typy TM proteinů.

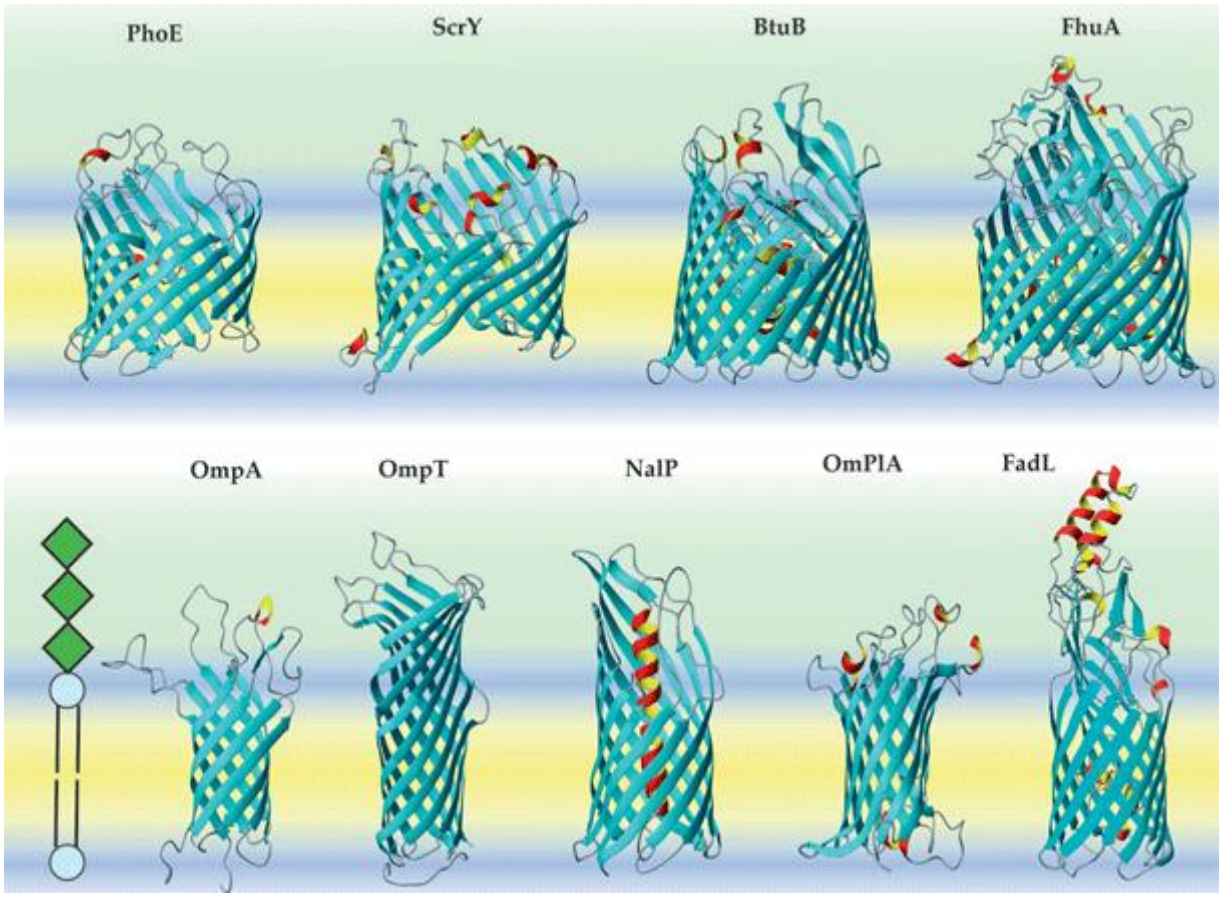
První jsou složeny ze svazku TM alfa helixů (viz Obr. 1) a mají klíčovou roli v biologických procesech, jako je komunikace s prostředím, přenos signálů a transport látek (Almen et al., 2009). Důležitost TM proteinů podtrhuje odhad, že lidský genom pro ně obsahuje 5539 genů, což je přibližně 26% strukturních genů (Fagerberg et al., 2010). Bohužel přes značné úsilí zůstává trojrozměrná struktura většiny těchto proteinů neznámá (Hopf et al., 2012), a tím roste důležitost predikčních metod. V této práci je detailně popsán postup predikce helikálních TM proteinů pomocí programů Rosetta (Yarov-Yarovoy et al., 2006) a EVfold_membrane (Hopf et al., 2012).



Obrázek 1: Ukázka TM proteinů ze svazku alfa helixů (převzato z Hopf et al., 2012)

Druhým typem TM proteinů v této práci jsou beta barely, které jsou k nalezení především ve vnější membráně Gram negativních bakterií, mitochondrií a chloroplastů (Waldispuehl et al., 2008). U mitochondrií a chloroplastů hrají kritikou roli při přenosu proteinů do těchto organel (Schleiff a Soll, 2005), u bakterií se účastní tvorby toxinových pórů (Delcour, 2002), a dokonce několik z nich i v enzymatické činnosti, jako například u *Escherichia coli* fosfolipáza OMPLA a proteáza OmpT (Bishop, 2008; Song et al., 1996). Beta barely jsou převážně tvořeny sérií antiparalelních beta řetězců, které tvoří válcovitou strukturu připomínající soudek (Schulz, 2000) (viz Obr. 2). U Gram negativních bakterií se odhaduje, že beta barely tvoří alespoň 3% všech

strukturních genů (Freeman a Wimley, 2010). V této práci je rozebráno, jak predikují strukturu beta barelových TM proteinů programy 3D-SPOT (Naveed et al., 2012) a EVfold_bb (Hayat et al., 2015).



Obrázek 2: Ukázka TM beta barelů z vnější membrány u Gram negativních bakterií (převzato z internetového zdroje [1])

[1] https://www.uni-kassel.de/fb10/institute/biologie/fachgebiete/biophysik/prof-dr-joerg-h-kleinschmidt/forschungsschwerpunkte/b-barrel.html

5 Predikce membránových proteinů – problematika

Znalost trojrozměrné struktury proteinu je extrémně důležitá pro plné pochopení jeho molekulární funkce a meziproteinové interakce (Marks et al., 2011). Existuje rozsáhlá databáze PDB (Protein Data Bank), která slouží jako primární archiv strukturních dat biologických makromolekul. Obsahuje více než 110 tisíc (duben 2016) makromolekulárních struktur a je široce používána predikčními metodami. Funkčnost PDB databáze popisuje Berman et al. (2000).

Potřeba znalosti struktury membránových proteinů je navíc podtržena odhadem, že 60% léků je zacílena právě na ně (Overington et al., 2006). Nicméně v PDB databázi je jich konstantně uloženo jen okolo 2,5% z celkového počtu (Tusnady et al., 2004).

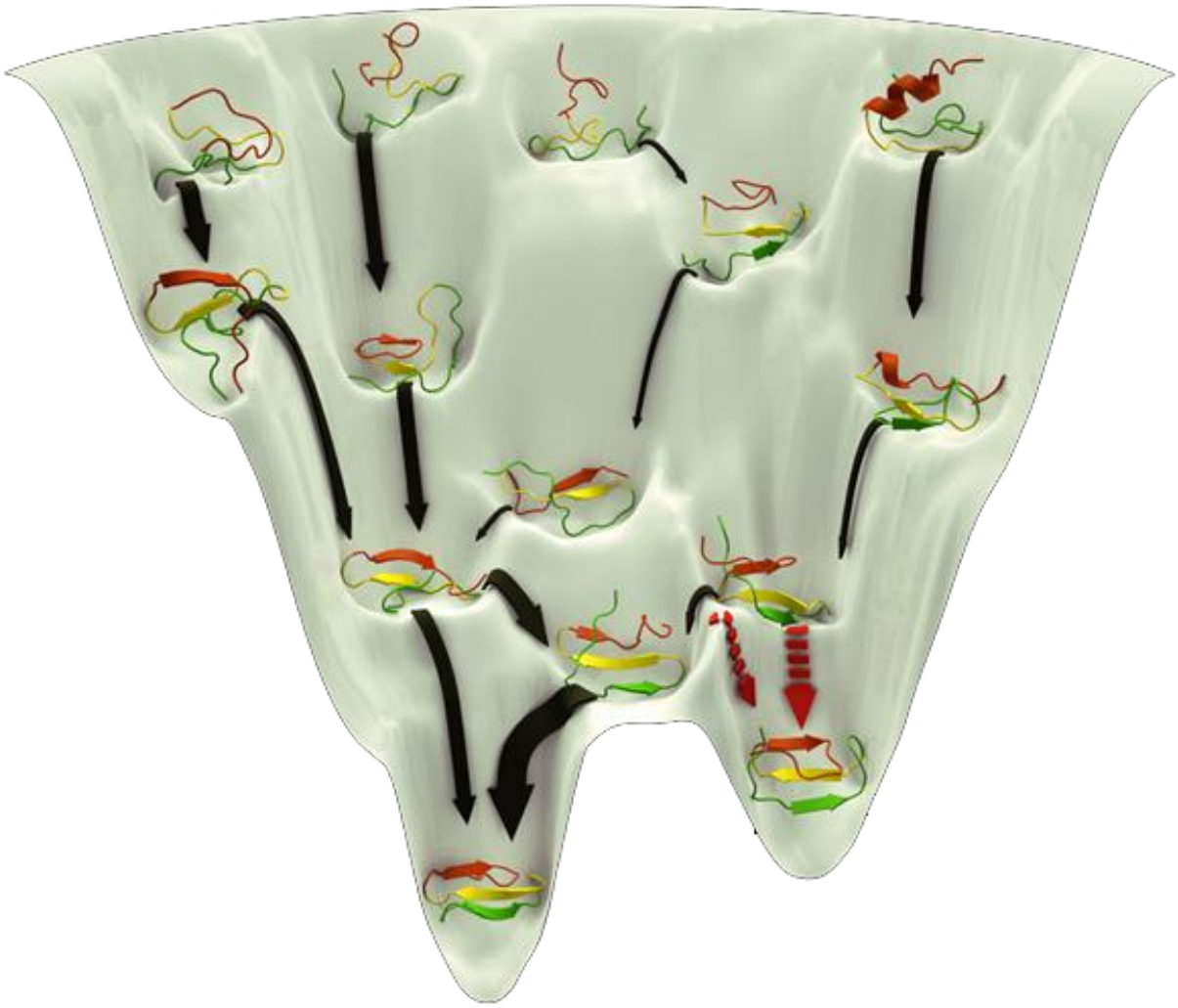
Proteinová struktura je typicky určována experimentálně pomocí X-ray krystalografie a NMR spektroskopie, bohužel membránové proteiny často znamenají pro tyto metody vážný problém (Bill et al., 2011). První je, že se membránové proteiny často nevyskytují v membráně organismu v dostatečném množství, kvůli čemuž je často nutné tvořit rekombinační proteiny pomocí expresních systémů (Bill et al., 2011). Další problém spočívá v purifikaci proteinů ve stabilní formě, protože pro tvorbu krystalů je nejdříve nutné proteiny vyjmout z membrány pomocí detergentů, které ho mohou poškodit, či změnit nativní konformaci, což je přirozená 3D struktura proteinu (Bill et al., 2011).

Další možností je využití molekulární dynamiky, která je převážně používána pro simulaci mezi-proteinové interakce a interakce proteinu s membránou (Lindahl a Sansom, 2008). Tento způsob je ale nesmírně výpočetně náročný. Nicméně již byly podniknuty experimenty využití molekulární dynamiky pro simulaci sbalování proteinů, jako například WW domény, což je malá proteinová doména, která se váže na sekvence bohaté na prolin (Shaw et al., 2010).

Třetí možností je výpočetní proteinová predikce, kterou lze obecně rozdělit do dvou kategorií. U jednodušší se jedná o takzvané homologní nebo templátové modelování, kdy k danému predikovanému proteinu existuje alespoň jeden již experimentálně zjištěný homologní protein, který je pak použit jako šablona (Baker a Sali, 2001). Druhá kategorie je volné modelování, kdy není nalezený žádný použitelný protein pro homologní modelování (Baker a Sali, 2001). Tato kategorie je označována jako *ab initio* nebo de novo predikce a je tématem této práce. Důležitost *ab initio* predikce navíc podtrhuje zjištění, že homologní modelování dokáže pokrýt jen 10% lidských TM proteinů (Hopf et al., 2012).

Výpočetní predikce jsou obecně založeny na předpokladu, že přirozená konformace proteinu má v daném prostředí nejméně energie. To znamená, že celá sekvence aminokyselin je sbalená tak, aby postranní řetězce a peptidická kostra byly uspořádány co nejvýhodněji. Například, aby hydrofobní aminokyseliny nebyly vystaveny vodnímu prostředí (Baker a Sali, 2001). S tím je spojeno prozkoumání ohromného množství nastavení všech možných úhlů v peptidové kostře i postranních řetězcích samotných aminokyselin, čemuž se v podstatě věnuje molekulární dynamika. Jak již bylo napsáno, tento proces je nesmírně náročný a na dnešních počítačích prakticky nemožný, proto si každá

metoda věnující se výpočetní predikci snaží nějak vypomoci (Baker a Sali, 2001; Durham et al., 2009). To znamená krokově nasimulovat kontinuální sbalování proteinu v prostředí s pomocí experimentálně zjištěných konzervovaných hodnot úhlů nebo kontaktů vzdálených aminokyselin a vytvoření metody, která bude schopná rozeznávat energeticky výhodnější konformace. Přiblížení takové simulace zobrazuje obrázek 3.



Obrázek 3: Diagram zobrazující předpoklad, že protein je sbalován pomocí energetické minimalizace. Vysokoenergetické konformace na začátku diagramu (nahoře) se postupně mění a snižují energii, až do energetického minima, kterým by měla být přirozená konformace. Čárkované červené šipky zobrazují nesprávné sbalování do lokálního energetického minima, kde může protein skončit in silico i in vivo (převzato z internetového zdroje [2])

Metody pro predikci 3D struktury obvykle používají velké množství externích pomocných programů pro dílčí problémy. Například pro predikci helikálních membránových proteinů je esenciální predikce úseků TM helixů v sekvenci. K tomuto účelu jsou používány speciálně zaměřené programy,

[2] https://www.behance.net/gallery/10952399/Protein-Folding-Funnel

kdy 12 nejlepších bylo porovnáno v článku Reeb et al. (2015). Moje práce na tyto externí programy většinou jen odkazuje a zabývá se hlavně metodikou, která je použita pro tvorbu 3D modelu. Existuje recentní review (Leman et al., 2015), které poskytuje obecný pohled na celkovou problematiku membránového modelování.

5.1 CASP – posouzení predikce proteinových struktur

The Critical Assessment of protein Structure Prediction (CASP) je v podstatě soutěž, která vznikla v roce 1994 a koná se každé dva roky. Jejím hlavním cílem je objektivně posoudit současné schopnosti v oblasti predikce struktury proteinů. V roce 2014 se konal CASP11, kterého se zúčastnilo 208 výzkumných skupin a z toho 85 automatických serverů, které mohou být volně přístupné na webu. Účastníci mají za úkol predikovat modely již experimentálně zjištěných (nejčastěji pomocí NMR nebo rentgenové krystalografie), ale nezveřejněných trojrozměrných struktur proteinů (Moult et al., 2014). Tyto proteiny jsou rozděleny do tří kategorií podle způsobu predikce, *ab initio*, homologní modelování a v CASP 10 nově zavedená kategorie „contact-assisted“, kdy algoritmy mohou využít experimentálně zjištěná data týkající se predikovaného proteinu jako data z chemického posunu NMR (Taylor et al., 2014).

Výsledné modely jsou hodnoceny podle mnoha kritérií, aby se co nejlépe definovala slabá a silná místa současné predikce. Z těch novějších je možné uvést například schopnost vyhlazování modelů (Rafinace modelu). Jedná se o metodu, která podrobí již vytvořený model výpočetně náročnému algoritmu, který se snaží optimalizovat jeho konformaci pomocí malých pohybů torzních úhlů s ohledem na všechny atomy v proteinu (Nugent et al., 2014). Dále se vyhodnocuje schopnost identifikace a predikce neuspořádaných oblastí proteinu, které v nativní konformaci nemají stabilní 3D strukturu. Jedná se především o regulační a vazebné oblasti proteinu (Monastyrskyy et al., 2014b). Predikci modelu může také velmi usnadnit schopnost předpovědět pravděpodobné kontakty aminokyselin v 3D konformaci, které jsou jinak v sekvenci od sebe vzdálené (Monastyrskyy et al., 2014a). Speciálně se pak vyhodnocuje i odhad kvality modelů. Jedná se o odhad dané výzkumné skupiny, jak moc je její vytvořený model spolehlivý (Kryshtafovych et al., 2014). Spolehlivost se často určuje podle množství a kvality vstupních dat, které jsou potřeba pro predikci daného modelu.

5.2 RMSD – měřítko úspěšnosti predikce

Model poskytnutý predikční metodou je potřeba nějak porovnat s experimentálně zjištěnou nativní (cílovou) konformací, aby bylo možné posoudit jejich shodu. K tomuto účelu se v celé bioinformatice široce používá parametr Root-mean-square distance (RMSD), definovaný jako

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N}\left(\left(x_i^a - x_i^b\right)^2 + \left(y_i^a - y_i^b\right)^2 + \left(z_i^a - z_i^b\right)^2\right)}{N}}$$

,

kde se měří odchylka vzdáleností N párů atomů v Å. V porovnávání se nejčastěji používají C_α atomy, části sekundárních struktur, nebo kompletní analýzy přes všechny atomy v závislosti na požadovaném výstupu. Pro správný výpočet je nutné nejdříve oba porovnávané proteiny přeložit přes sebe tak, aby RMSD bylo co nejmenší (Coutsias et al., 2004) (viz Obr. 4). K tomu se používá velmi rozšířený Kabschův algoritmus (Kabsch, 1976), nebo novější metoda za použití kvaternionů, což jsou uspořádané čtveřice reálných čísel se speciálně definovanými výpočetními operacemi (Coutsias et al., 2004).

Obrázek 4: Ukázka překrytí predikovaného modelu u třech různých nativních struktur (převzato z Barth et al., 2009)

Samotné RMSD může být zavádějící při porovnávání úspěšnosti metody u různě velikých proteinů. Pokud například nějaká metoda dokáže vytvořit model proteinu o velikosti 100 aminokyselin s RMSD 4 Å vůči nativní struktuře, tak je to určitě menší úspěch, než model proteinu o délce 400 se stejnou hodnotou RMSD (Irving et al., 2001). Jistou formu normalizace zavádí RMSD100, které je definované jako

$$rmsd_{100} = \frac{rmsd}{1 + \ln\sqrt{\frac{N}{100}}},$$

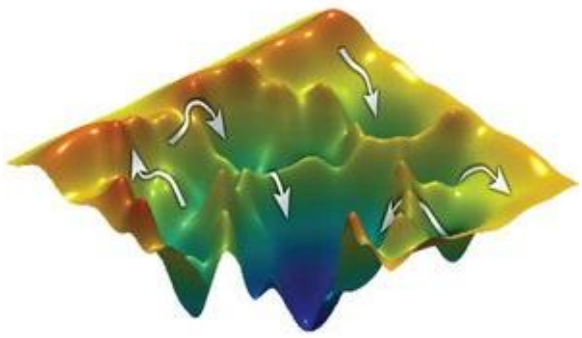
kde N je počet aminokyselin v proteinu a 100 je počet aminokyselin, přes který probíhá normalizace (Carugo a Pongor, 2001). Konstanta 100 byla vybrána jako průměrný počet aminokyselin v proteinové doméně (Xu a Nussinov, 1998).

5.3 Hledání globálního energetického minima - simulated annealing

Simulované žíhání (simulované ochlazování, simulated annealing) je algoritmus založený na hledání globálního optima pro daný systém. Tento systém se skládá ze stavů, do kterých může systém přejít a v rámci nich je i stav globálního optima. Pokud systém přejde z jednoho stavu do druhého, tak je vyhodnoceno, jestli v systému došlo ke zlepšení nebo zhoršení směrem k optimu.

Aby algoritmus našel globální optimální stav, musí občas přejít do stavu horšího, aby se vymanil ze špatného lokálního optima (energetického minima). Jak moc může přecházet do horších stavů, určuje proměnná „teplota“, která je iterativně zmenšována, a tím se zmenšuje možnost přechodu do horšího stavu (Kirkpatrick et al., 1983). Tímto postupem je dané, kolik kroků algoritmus vykoná, ale není znám jeho výsledek. Algoritmu se tím řadí do skupiny Monte Carlo algoritmů. Navíc se používá Metropolisův algoritmus, kdy každý další stav při annealingu je navíc ovlivněn jeho pravděpodobností vůči předcházejícímu (Metropolis et al., 1953).

V predikci proteinů se tyto algoritmy používají na simulaci sbalování, kde stavy reflektují jednotlivé konformace proteinu. V prvních fázích jsou možné velké konformační změny a s klesající teplotou se postupně pohyby modelu omezují (Karakas et al., 2012; Simons et al., 1997) (viz Obr. 5). Tím dochází k hledání globálního energetického minima. Metropolisův algoritmus je pak používán na přijetí nebo odmítnutí konformační změny, to znamená přechodů do jednotlivých stavů (Simons et al., 1997).



Obrázek 5: Ukázka schopnosti simulovaného žíhání vyhnout se nesprávným lokálním energetickým minimům, které mají barvu zelenou. Globální energetické minimum je obarveno modře (převzato z internetového zdroje [3]).

5.4 Hodnocení modelů – energetická funkce

Správné ohodnocení modelů je esenciální pro každý algoritmus zabývající se trojrozměrnou predikcí proteinů (Simons et al., 1997), ať už je potřeba ohodnotit výsledný model, nebo dílčí konformační změny během simulace. To se provádí pomocí hodnotící (energetické) funkce s cílem rozlišit dobře sbalené modely (méně než 4 Å RMSD) od nesprávně sbalených (Simons et al., 1999b). Nejedná se o skutečnou energii (kJ/mol), ale o soubor výrazů hodnotící jednotlivé aspekty modelu (např. kompaktnost modelu, párování typů aminokyselin, správné rozmístění hydrofobních a hydrofilních postranních řetězců, správné párování beta řetězců, atd.) (Rohl et al., 2004). Podstatou funkce je čistě statistická informace, což znamená, že pokud se například nějaká aminokyselina

[3] http://bioforces.blogspot.cz/2011/10/single-molecule-fluorescence.html

vyskytuje v přírodě s určitou pravděpodobností uvnitř proteinu, a v modelu je na povrchu, bude hodnocení daného výrazu o tuto pravděpodobnost zhoršené (Simons et al., 1999b).

Během simulace se často používají jen jednotlivé části dané energetické funkce a další se časem přidávají. Tento postup je z důvodu rostoucí složitosti konformace modelu, a tudíž je potřeba i komplexnější hodnocení (Karakas et al., 2012; Simons et al., 1997). Pokud by se celková funkce používala během celé simulace, tak by vzhledem k náročnosti hodnocení došlo ke značnému časovému nárůstu.

Pokud by funkce byla dokonalá, tak by stačilo vygenerovat velké množství modelů, a jako ten nejlepší vzít ten s nejlepším hodnocením. Bohužel to ale tak není, a proto je nutné k výběru použít i jiné metody, jako například klastrování.

5.5 Contact order – ovlivnění rychlosti sbalování

Contact order (CO) nebo též relativní contact order poskytuje informaci o rozmístění aminokyselin v dané konformaci. Jedná se o průměr vzdáleností v primární sekvenci mezi jednotlivými definovanými páry aminokyselin, které jsou spolu v kontaktu v trojrozměrné struktuře. Studie ukázaly, že se zvyšujícím se CO klesá rychlost sbalování proteinů a je v tomto ohledu rozhodujícím faktorem (Grantcharova et al., 2001; Plaxco et al., 1998). V predikci se používá například pro porovnání CO ve vytvořených modelech s experimentálně zjištěnými strukturami o stejné velikosti (Bonneau et al., 2002). Obecně je CO definovaný jako

$$CO = \frac{1}{L \cdot N} \sum^{N} \Delta S_{i,j}$$

,

kde N je počet kontaktů, L je celkový počet aminokyselin v proteinu a $\Delta S_{i,j}$ udává délku sekvence mezi aminokyselinami tvořící kontakt. Definice kontaktu může být různá, například je jako kontakt brán každý pár aminokyselin, které mají u sebe atomy (kromě vodíku) blíž než 6 Å (Plaxco et al., 1998), nebo dvojice aminokyselin, které mají C_β atomy ve vzdálenosti do 8 Å a jsou sekvenčně vzdáleny alespoň o 3 aminokyseliny (Bonneau et al., 2002). V některých analýzách se využívá i absolutní CO, které není normalizované přes L (Bonneau et al., 2002).

6 Rosetta

6.1 Úvod

Rosetta je komplexní nástroj pro modelování a analýzu proteinových a nukleotidových struktur. Vyvinul ji Dr. David Baker z univerzity ve Washingtonu jako nástroj pro doplnění molekulárních a bio-fyzikálních studií sbalování malých solubilních proteinů (Simons et al., 1997). Postupně přesáhla univerzitní prostředí a pracuje na ní mnoho institucí, které se sdružují do takzvaných členů RosettaCommons. Samotný software se skládá ze souboru oddělených algoritmů (protokolů), které spolu ale vzájemně spolupracují. V této práci nás hlavně zajímají protokoly pro *ab initio* modelování a hlavně „membrane *ab initio*“ protokol, který byl vytvořen pro predikci helikálních membránových proteinů.

6.2 *Ab initio* protokol

Tento protokol byl vyvinut pro vytváření modelů malých solubilních proteinů z pouhé sekvence aminokyselin, kdy není dostupný žádný homologní protein, který by mohl sloužit jako šablona během modelování. Základní myšlenka algoritmu je použití informace z mnoha již experimentálně definovaných trojrozměrných struktur proteinů pomocí velmi krátkých úseků s nejlepší profilovou podobností (Simons et al., 1997). Profilová podobnost je vysvětlena pomocí obrázku 6.

Obrázek 6: Obrázek ukazuje nalezení nejlepší profilové shody pro horní sekvenci. Kdyby se porovnávala jen samotná shoda sekvence, tak by sekvence 1 a 2 byly na stejné úrovni. Ale v profilovém porovnání hraje důležitou roli i informace ze sekvence 3, která má na druhé pozici Alanin, což rozhodne ve prospěch sekvence 2.

Před predikcí trojrozměrného modelu je vytvořen soubor obsahující veškeré možné úseky devíti a tří aminokyselin z predikované sekvence. K těmto jednotlivým úsekům jsou dohledány profilově podobné úseky z proteinů, které již mají nalezenou trojrozměrnou strukturu, a jsou zaznamenány jejich torzní úhly. Tyto dohledané úseky jsou označovány jako fragmenty (Simons et al., 1997). V roce 2004 v CASP6 (Bradley et al., 2005) byly k nalezení těchto úseků použity programy Pfam4 (Bateman et al., 2004) a PSI-BLAST (Altschul et al., 1997). Výběr fragmentů je dále ovlivněn předpovězením sekundární struktury v daném predikovaném úseku (Simons et al., 1999a), kdy souhlas

v sekundární struktuře zlepšuje hodnocení fragmentu. V CASP6 byly k predikci sekundárních struktur použity programy PSIPRED (Jones, 1999), Sam-T99 (Karplus et al., 1999) a JUFO (Meiler et al., 2002). Pro každý úsek devíti a tří aminokyselin je takto vybráno 25 nejlepších fragmentů (Simons et al., 1997).

Pro zjednodušení predikce jsou postranní řetězce aminokyselin nahrazeny pseudo-atomem, který je označován jako těžiště (Simons et al., 1999b). Ten je definovaný jako průměrná pozice všech atomů postranního řetězce přes všechny rotamery v PDB databázi (Gray et al., 2003), kdy rotamer označuje jeho možnou konformaci (Dunbrack a Cohen, 1997).

Simulace sbalování proteinu začíná na plně nataženém řetězci, kde je náhodně vybráno místo, kde dojde k iniciaci sbalování nahrazením torzních úhlů řetězce z náhodně vybraného fragmentu pro dané místo o délce devíti aminokyselin. Tento proces následně pokračuje na náhodných místech řetězce, dokud není každý torzní úhel nahrazen alespoň jednou. Na začátku simulace způsobí vložení fragmentu velkou konformační změnu, ale jak se model stává více sbalený, tak vložení způsobí nejen změnu torzních úhlů v sekvenci fragmentu, ale způsobí změny i v blízkém okolí tak, aby bylo vložení slučitelné s aktuální konformací. Pohyb navíc musí být vykonán tak, aby se žádné dva atomy k sobě nepřiblížily více než 2,5 Å (Simons et al., 1997).

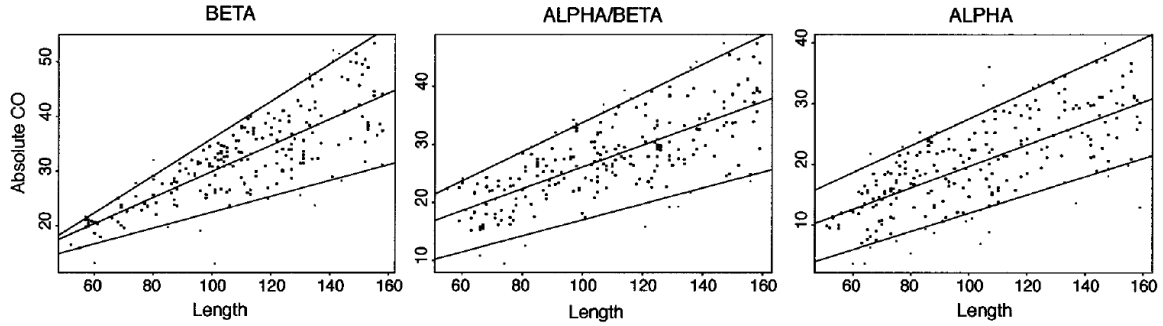
Simulace je rozdělena do několika fází, kdy v každé následující fázi dochází ke komplexnějšímu ohodnocení modelu po vložení fragmentu. Například v první fázi dochází jen k ohodnocení sterických překryvů a v další se již navíc hodnotí i kompaktnost modelu (Rohl et al., 2004). V jedné z posledních fází dojde k vyhlazení modelu pomocí fragmentů tří aminokyselin (Bradley et al., 2003), kdy pomocí metody Gunn (Gunn, 1998) dojde k úpravě konformace i v okolí fragmentu (Bonneau et al., 2001b).

Pokud jsou v sekvenci identifikovány jasně oddělené domény, jsou rozštěpeny a predikovány zvlášť výše popsaným algoritmem. Následně jsou pak spojeny do jednoho modelu (Bradley et al., 2003). Výslednému modelu je pak přiřazeno číslo energetickou funkcí (score), které simuluje jeho energii, to znamená, čím nižší hodnota funkce, tím lepší model (viz kapitola 6.5).

Jelikož simulované žíhání spadá do Monte Carlo algoritmů, je zřejmé, že výsledný model nemusí odpovídat přirozené konformaci proteinu vyskytující se v přírodě. Je to dáno náhodností vkládání fragmentů a hlavně krokově omezeným simulovaným žíháním. Prakticky většina vytvořených modelů je nesprávně sbalená a je nutné nalézt ten správný. Hodnocení modelů v tomto případě nestačí z důvodů nedokonalosti energetické funkce (Tsai et al., 2003). Tento problém je řešen generováním velkého množství modelů (čím víc, tím lépe), které následně vstupují do fáze filtrovací a klastrovací (Bonneau et al., 2002).

Filtrovací fáze má za úkol vyřadit evidentně špatně sbalené modely, které obsahují například nesprávně vytvořené beta listy (Bonneau et al., 2001b), nebo hodnotu CO, která nekoresponduje s očekávanou hodnotou pro proteiny podobné délky a obsahující podobné sekundární struktury (Bonneau et al., 2001b). Algoritmus používaný Rosettou, který má tendenci tvořit hlavně struktury

tvořené lokálními aminokyselinami, vykazuje nadbytečnou tvorbu modelů s nízkým CO (Bonneau et al., 2002). Byl tedy vytvořen přístup, který filtruje modely před vstupem do klastrování s ohledem na CO. Tento přístup je postavený na definování percentilů absolutního CO a délky proteinu pomocí reprezentativního souboru zjištěných proteinů. Tyto percentily jsou definovány pro tři kategorie proteinů podle výskytu sekundárních struktur na α, αβ a β (Bonneau et al., 2002) (viz Obr. 7).

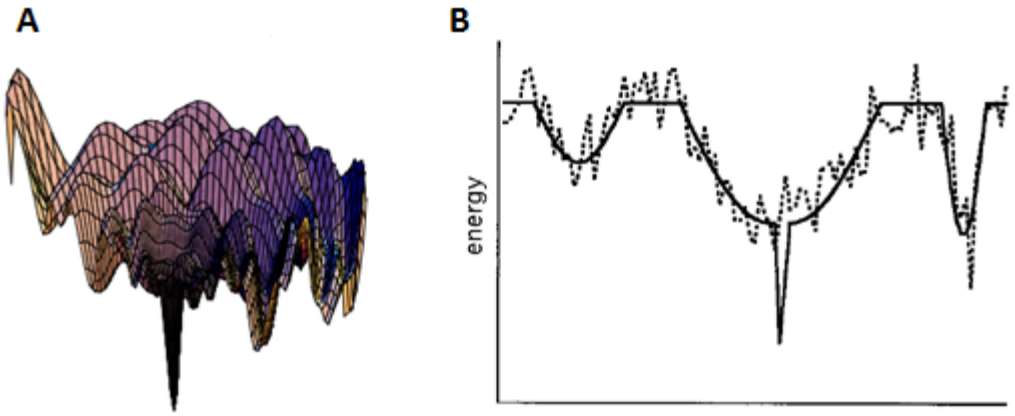


Obrázek 7: Definice percentilů na základě reprezentativního souboru proteinů. Proteiny jsou rozděleny do tří kategorií podle výskytu sekundárních struktur. Každý protein je pak umístěn do grafu podle své délky a absolutního CO. Graf je následně rozdělen na čtyři části, kdy čáry oddělují jednotlivé percentily (5%, 45%, 45%, 5%) výskytu proteinů (převzato z Bonneau et al., 2002)

Pokud tedy máme soubor modelů o nějaké velikosti, tak na základě třídy, délky a absolutního CO dojde k rozdělení souboru do grafů z obrázku 7. To způsobí, že předem definované percentily rozdělí soubor modelů a z každého percentilu je následně vzato odpovídající procento modelů s nejlepším score (5%, 45%, 45%, 5%) (Bonneau et al., 2002; Tsai et al., 2003). Nicméně nejjednodušší filtr má za cíl poskytnout klastrovací fázi jen určité procento modelů s nejnižším score (Shortle et al., 1998).

Fáze klastrování začíná vzájemným výpočtem C_α RMSD pro všechny vložené modely a každému je přiřazen počet podobných modelů na základě uživatelsky definovaného RMSD (např. menší než 8 Å (Bonneau et al., 2001a)). Dále je uživatelsky definovaná maximální velikost klastru (např. 100 modelů (Bonneau et al., 2001b)). Pokud model s největším množstvím podobných jich má více, než je povolená velikost klastru, tak dojde k iterativnímu zmenšování rozdílového RMSD, dokud není dosažena maximální velikost, nebo nedojde ke zmenšení rozdílového RMSD na hodnotu 3 Å. Následně dojde k zaznamenání modelu jako centrum klastru a jeho rozdílové RMSD a všechny proteiny v klastru jsou vyřazeny z populace a proces začíná znova, dokud jsou tvořeny klastry o určité velikosti (Bonneau et al., 2001a; Bonneau et al., 2001b; Shortle et al., 1998). Centra největších klastrů jsou pak předložena jako možné výsledky predikce (Shortle et al., 1998).

Klastrování je založeno na předpokladu, že přirozená struktura je energeticky nejvýhodnější a nachází se uprostřed nejširšího energetického trychtýře v energetické krajině, a že hodnocení založené převážně na hydrofobních interakcích dokáže tento trychtýř rozeznat (viz Obr. 8) (Shortle et al., 1998).

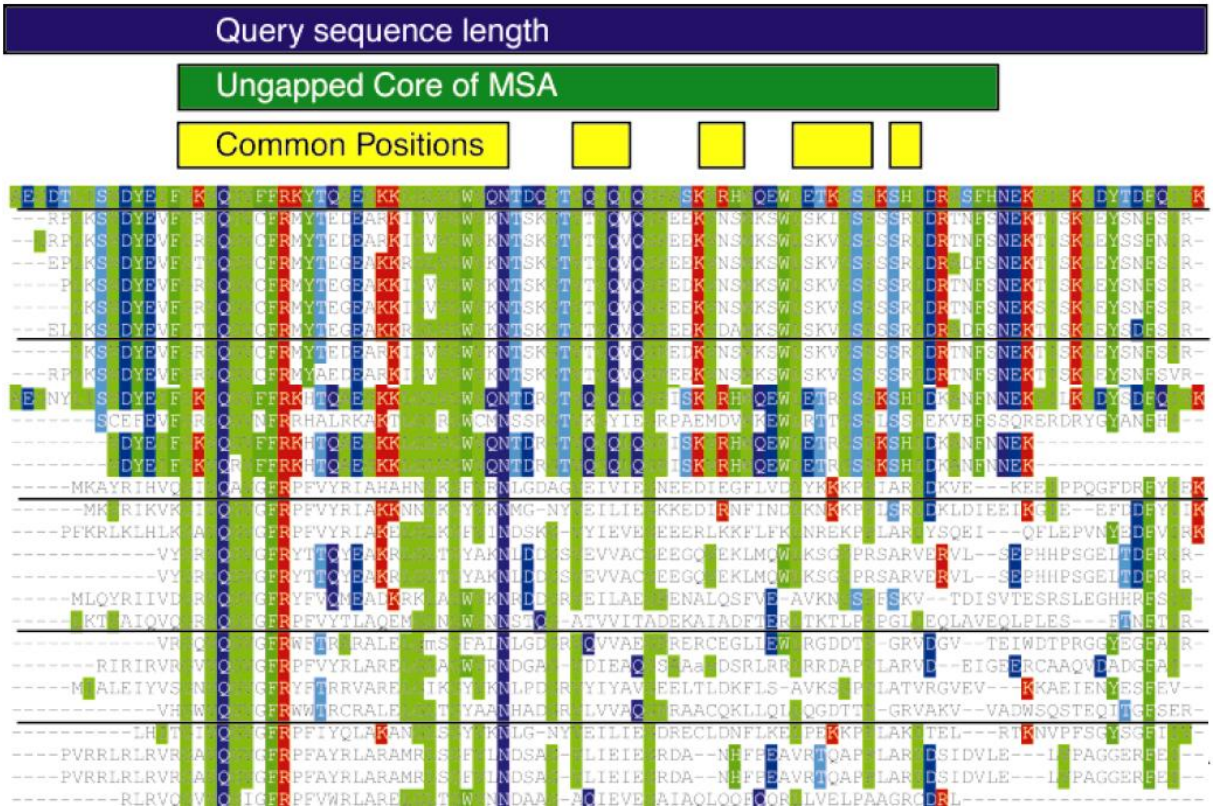


Obrázek 8: (A) Zobrazení různě širokých oblastí, kdy nejširší oblast vede k úzké prohlubni přirozené konformace (převzato z Bryngelson et al., 1995). (B) Ukázka propojení energie konformace (plná čára) s energetickou funkcí (přerušovaná čára), která ukazuje, že funkce nemusí nalézt nejnižší energii, ale dokáže identifikovat oblast, která k ní vede (převzato z Shortle et al., 1998).

Pro lepší výsledky je často mezi filtrováním a klastrováním zařazeno doplnění postranních řetězců aminokyselin, kdy těžiště jsou nahrazena rotamery postranních řetězců ze známých struktur, tak aby nedocházelo ke sterickým překryvům (Bonneau et al., 2001b; Kuhlman a Baker, 2000). Po každém nahrazení dojde k náhodným pohybům peptidické kostry okolo aminokyseliny a zjišťuje se energetická výhodnost vkládaného rotametru (Misura a Baker, 2005). Tento krok je ale velmi výpočetně náročný, takže se někdy nechává až na výsledky klastrování.

6.3 Využití informace z profilově podobných sekvencí

K lepšímu rozpoznání přirozené konformace je možné využít informaci z ostatních proteinů (Bonneau et al., 2001a). Vychází se z empirických pozorování, že pokud dvě sekvence aminokyselin sdílejí víc jak 25% sekvenční shodu a jsou delší než 60 aminokyselin, tak sdílejí téměř i stejnou konformaci (Abagyan a Batalov, 1997; Brenner et al., 1998; Sander a Schneider, 1991). Pomocí PSI-BLAST (Altschul et al., 1997) je vytvořen soubor sekvencí s alespoň 60% délkou predikované sekvence a podobnost je vyžadována od 20% do 60%. Následně se definují úseky, které jsou do jisté míry všem společné a sekvenční jádro, které spojuje společné úseky. Z těchto profilově podobných sekvencí jsou následně vybrány ty, které mají vzájemnou podobnost menší než 60% (Bonneau et al., 2001a) (viz. Obr. 9)



Obrázek 9: Modrý pruh zobrazuje délku sekvence predikovaného proteinu s PDB kódem 2ACY, zelený pruh je sekvenční jádro a žluté jsou společné úseky. Podtržené sekvence značí vybrané sekvence (převzato z Bonneau et al., 2001a). Různé barvy aminokyselin sdružují ty se stejnými fyzikálně chemickými vlastnostmi důležité pro porovnání.

Pro každou takto vybranou sekvenci je standardním algoritmem Rosetty vytvořeno 1000 modelů pro celý řetězec a 1000 modelů jen pro sekvenční jádro. Tyto modely jsou následně klastrovány jen přes společné úseky.

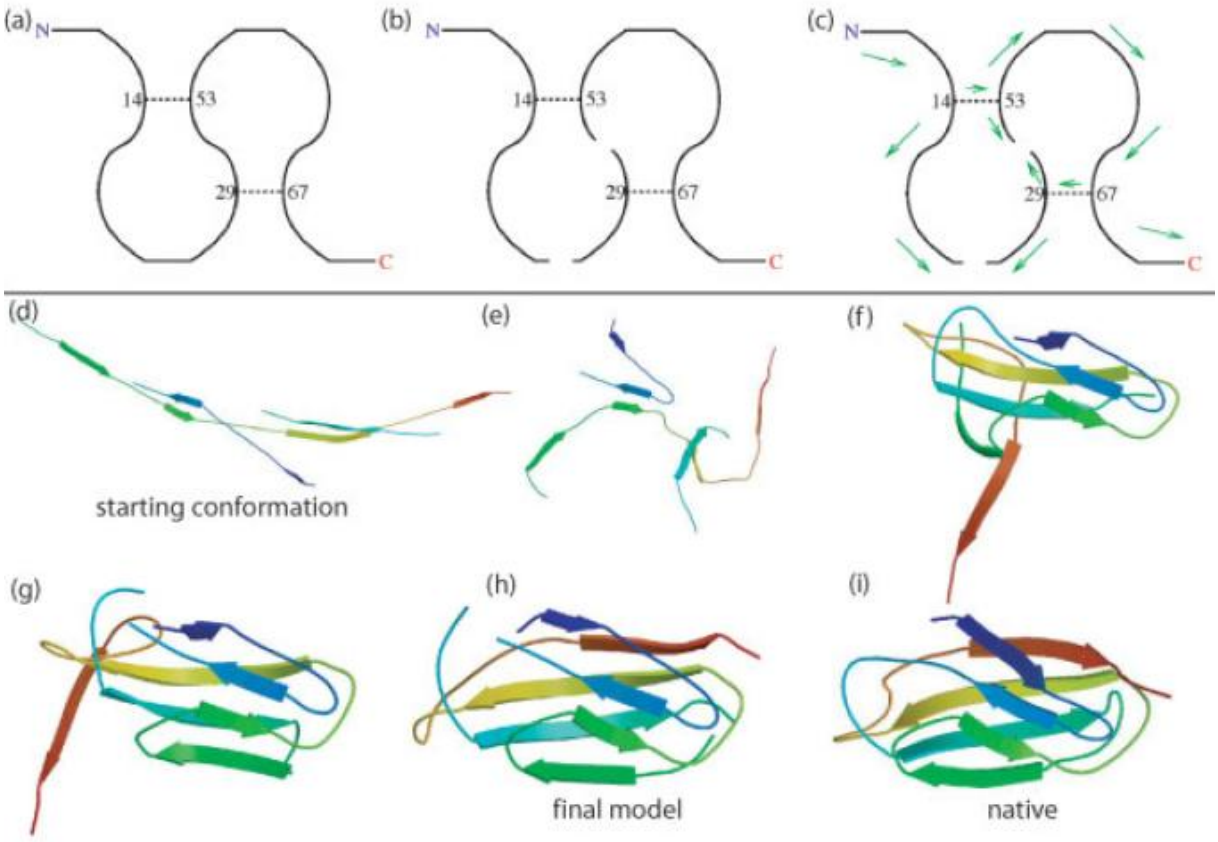
Následně jsou vytvořeny modely pro predikovanou sekvenci a standardně klastrovány. V ideálním případě by centrum klastru obsahující přirozenou konformaci mělo mít odpovídající klastry z dodatečných sekvencí porovnávané přes společné úseky.

6.4 Nelokální beta skládané listy

Jelikož algoritmus predikce je založen na lokálním nahrazování torzních úhlů, jsou přirozeně upřednostňovány lokální beta struktury, které se zformují jako první. Tím může dojít k nesprávnému sbalení lokálních beta struktur na úkor sekvenčně vzdálených beta struktur (Bradley a Baker, 2006).

Tento problém je řešen pomocí spojů mezi vzdálenými aminokyselinami, které jsou během standardního skládání udržovány u sebe. Tyto spoje jsou odvozeny od zjištěných proteinů, kdy jsou zaznamenány interagující aminokyseliny v beta listech a jejich paralelní či antiparalelní orientace. Pokud jsou tyto páry nalezeny na predikované sekvenci, kde je předpokládaná beta struktura, dojde ke spojení pomocí náhodně vybraného páru. Zároveň dojde k rozštěpení peptidového řetězce v predikovaných smyčkách mezi těmito sekvencemi (viz Obr. 10). Tím dojde k modifikaci

peptidového řetězce, schopného podstoupit standardní algoritmus. V pozdějších fázích simulace je do energetické funkce zavedeno hodnocení pro udržování rozštěpených konců u sebe (Bradley a Baker, 2006).



Obrázek 10: Konstrukce a použití vzdálených spojů (převzato z Bradley a Baker, 2006)

6.5 Hodnocení modelu

Energetická funkce Rosetty je založena na Bayesově větě, která udává podmíněnou pravděpodobnost nějakého jevu (Simons et al., 1997). Po upravení věty pro hodnocení pravděpodobnosti modelu vypadá takto

$$P(structure \mid sequence) = P(structure) \times \frac{P(sequence \mid structure)}{P(sequence)}$$

,

a je založena čistě na statistické informaci viz kapitola 6.5. Jednotlivé dílčí výrazy funkce jsou zobrazeny na obrázku 11. P(sekvence) je konstanta a je ve výpočtu zanedbána. Samotné score, kterým je model hodnocen, je -log P(structure | sequence). Použití -log udává, že čím menší score, tím by měl být model blíže přirozené konformaci, ale to nemusí vždy platit, a proto je například používáno klastrování.

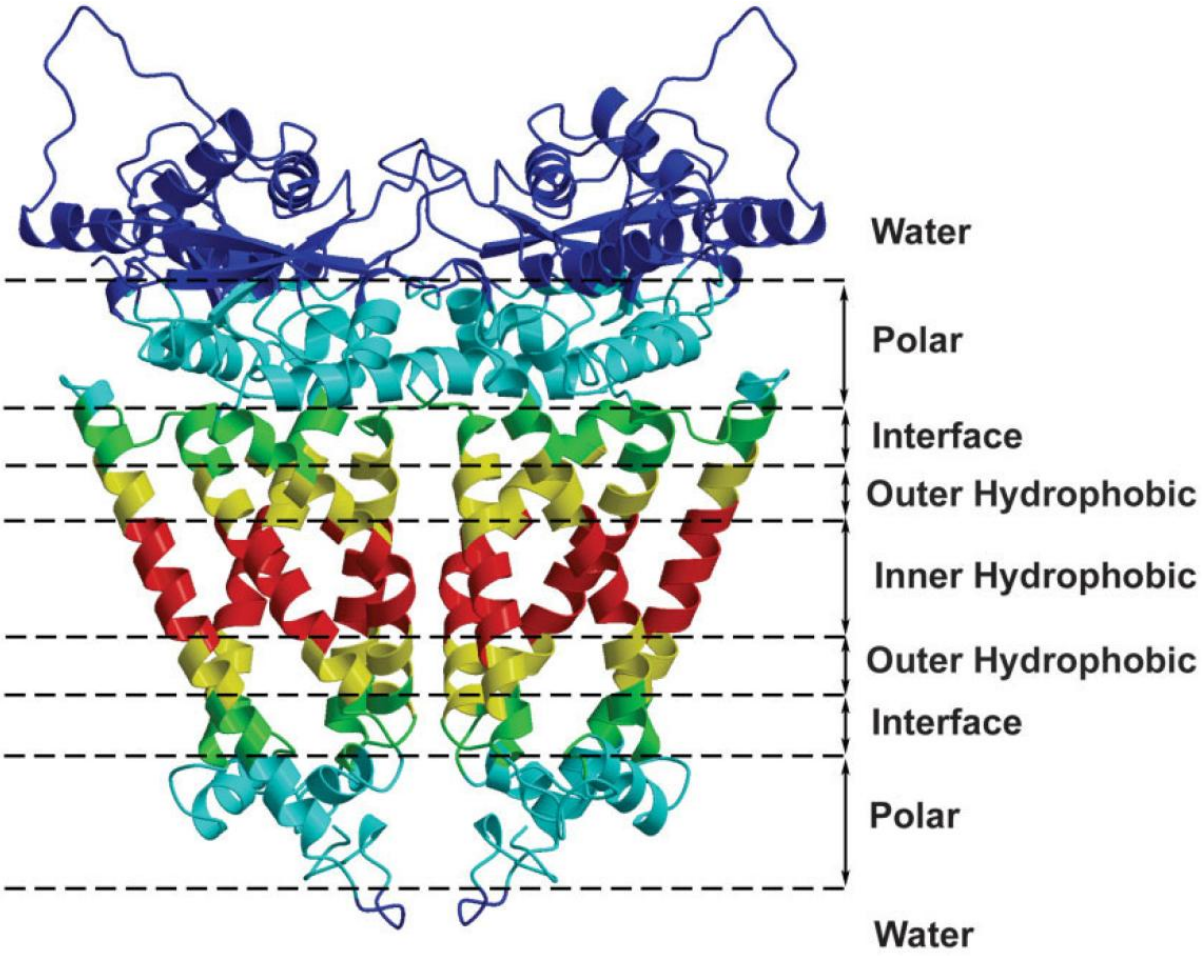
Probability density	Functional form	Putative physical origin
I. Sequence dependent	P(sequence|structure)	
A. Residue-environment	P_{env}	Hydrophobic effect
B. Residue-residue	P_{pair}	Electrostatics, disulfides
C. Local sequence-structure	ISITES	Sequence-local structure
D. Packing orientation	$P_{packing\text{-}seq}$	Packing geometry
II. Sequence independent	P(structure)	
A. Secondary structure packing	$P_{HH\text{-}\phi\theta}$	Helix-helix packing
	$P_{HH\text{-}dist}$	
	$P_{HS\text{-}\phi\theta}$	Helix-strand packing
	$P_{HS\text{-}dist}$	
	$P_{SS\text{-}\phi\theta}$	Strand-strand packing
	$P_{SS\text{-}dist}$	
B. Strand hydrogen bonding	P_{SShb}	Hydrogen bonding
C. Strand assembly in sheets	P_{sheet}	Hydrogen bonding
D. Hard sphere repulsion	VdW^b	Steric repulsion
E. Structure compactness	$P_{density}$	Hydrophobic effect,
	Radius of gyration	Van der Waals interactions
F. Local structure	$P_{local\text{-}struct}$	Local structure preferences
G. Packing orientation	$P_{packing\text{-}struct}$	Packing geometry

Obrázek 11: Přehled dílčích výrazů energetické funkce (převzato z Simons et al., 1999b)

6.6 Membrane *ab initio* protokol

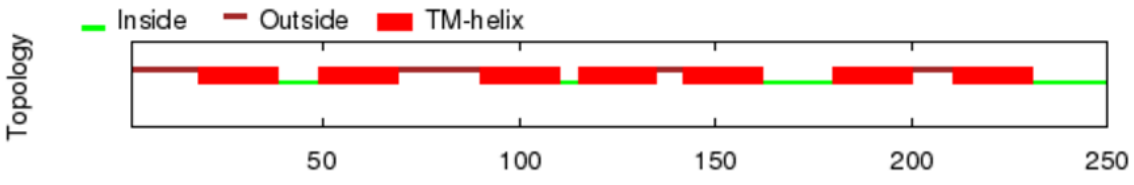
Počínaje analýzou fotosyntetického reakčního centra (Rees et al., 1989) bylo zjištěno, že velké hydrofobní aminokyseliny jako fenylalanin, leucin, isoleucin a valin jsou v TM helixech častěji vystaveny do membránového prostředí (Adamian et al., 2005; Ulmschneider et al., 2005). Naopak malé aminokyseliny jako glycin, alanin, serin a threonin mají tendenci vytvářet mezihelixové interakce (Eilers et al., 2002; Javadpour et al., 1999).

Tyto a další poznatky vedly k vytvoření nového protokolu pro predikci helikálních TM proteinů. Algoritmus založený na fragmentech a energetická funkce musely být adaptovány na anizotropické membránové prostředí (Yarov-Yarovoy et al., 2006). Kvůli tomu byl v solubilním prostředí vytvořen model membrány po vzoru White a Wimley (1999), kdy hydrofobní část byla navíc rozdělena na vnitřní a vnější (viz Obr. 12) (Yarov-Yarovoy et al., 2006).



Obrázek 12: Definice membrány (převzato z Yarov-Yarovoy et al., 2006)

Algoritmus začíná standardním nalezením fragmentů, ke kterému je použit program Sam-T99 (Karplus et al., 2001). Oproti solubilním proteinům je navíc důležité zjistit celkovou topologii proteinu kvůli správnému usazení v membráně. Pro zjištění topologie transmembránových úseků je používán program OCTOPUS (Viklund a Elofsson, 2008), a pomocí nich se zjišťuje, na které straně membrány jsou jednotlivé C- a N- konce helixů (viz Obr. 13) (Yarov-Yarovoy et al., 2006).



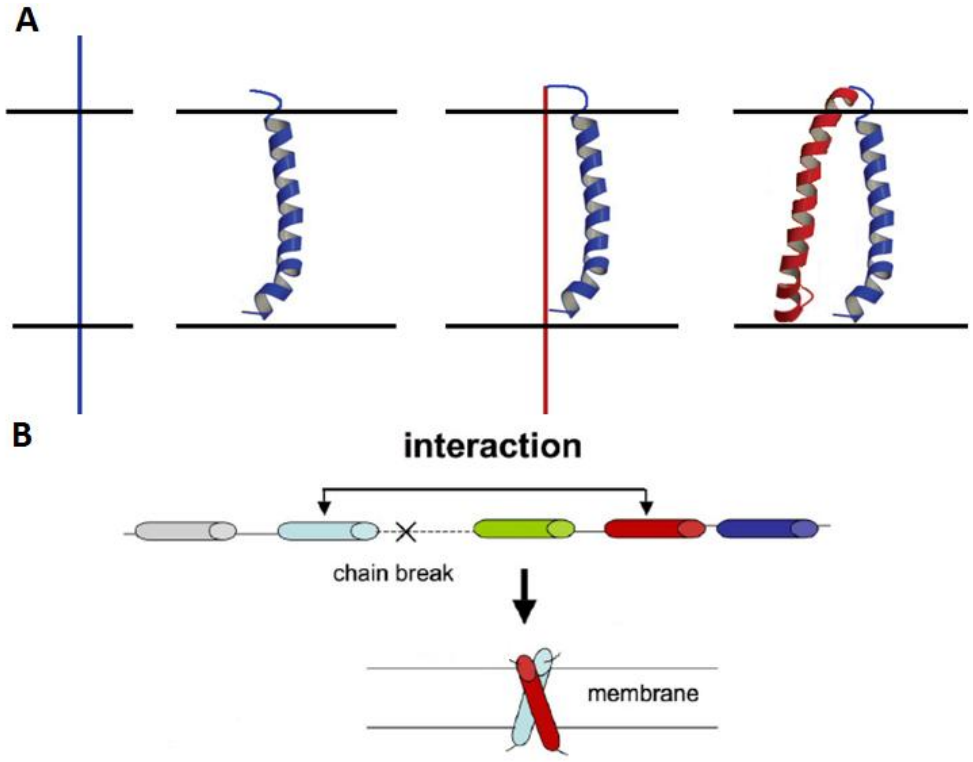

```
Sequence length: 250 aa.
Sequence:
MCCAALAPPMAATVGPESIWLWIGTIGMTLGTLYFVGRGRGVRDRKMQEFYIITIFITTI
AAAMYFAMATGFGVTEVMVGDEALTIYWARYADWLFTTPLLLLDLSLLAGANRNTIATLI
GLDVFMIGTGAIAALSSTPGTRIAWWAISTGALLALLYVLVGTLSENARNRAPEVASLFG
RLRNLVIALWFLYPVVWILGTEGTFGILPLYWETAAFMVLDLSAKVGFGVILLQSRSVLE
RVATPTAAPT
```

```
OCTOPUS predicted topology:
ooooooooooooooooMMMMMMMMMMMMMMMMMMMiiiiiiiiiiMMMMMMMMMMM
MMMMMMMMMoooooooooooooooooooooMMMMMMMMMMMMMMMMMMMMiiiiMMMMMM
MMMMMMMMMMMMMMMoooooooMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiM
MMMMMMMMMMMMMMMMMMMMooooooooooMMMMMMMMMMMMMMMMMMMMiiiiiiiii
Iiiiiiiii
```

Obrázek 13: Ukázka predikce topologie pro Bacteriorhodopsin. Každé aminokyselině je přiřazené písmeno, které identifikuje její pozici: o – vnější prostor, M – membrána a i – vnitřní prostor

Po vzoru nelokálních beta skládaných listů (viz kapitola 7.4) jsou dále dohledány kontakty mezi jednotlivými TM helixy (Barth et al., 2009) (viz kapitola 7.7). Dva TM helixy jsou pokládány za interagující, pokud je nalezeno alespoň 5 kontaktů s C_α atomy v okruhu 8 Å.

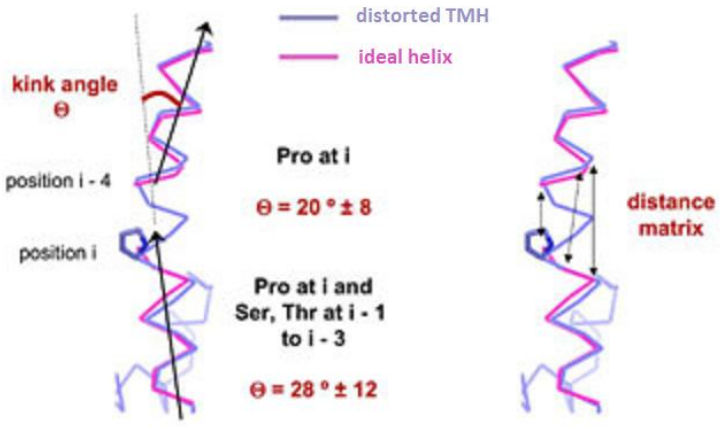
Simulace sbalování začíná tvorbou membrány, kdy jsou přes sebe přeloženy natažené predikované sekvence TM helixů a je vypočítána průměrná 2D pozice C_α atomů na obou stranách predikované membrány. Uprostřed těchto průměrných pozic je označen střed membrány a přímka jimi definovaná je brána jako její normála. Plochy membrány jsou pak vytvořeny jako kolmé roviny vzdálené od centra 30 Å. Do takto definované membrány je vložena jedna natažená sekvence jednoho vybraného helixu zprostředka proteinu (Yarov-Yarovoy et al., 2006). Pokud byly definovány interakce mezi helixy (nalezeny kontakty), tak je náhodně vybrán jeden z párů helixů a od něj jeden z kontaktů. V tomto případě dojde na počátku simulace k vložení těchto dvou helixů v podobě natažených sekvencí namísto jednoho, a tyto helixy jsou spojené vybraným kontaktem. V náhodném místě mezi helixy je sekvence rozštěpena v místě předpokládané smyčky a kontakt je udržován během simulace (Barth et al., 2009). Takto vložené helixy jsou podrobeny standardnímu vkládání fragmentů. Po určitém množství vložení je náhodně vybrán volný C nebo N konec vložených helixů a je připojen další helix, který je následně opět podroben vkládání fragmentů, dokud takto nejsou zpracovány všechny helixy. Helixy se mohou v membráně pohybovat, aby co nejlépe protínaly membránovou dvouvrstvu (Yarov-Yarovoy et al., 2006). Počátek simulace zobrazuje obrázek 14.



Obrázek 14: (A) Ukázka počátku simulace bez nalezeného kontaktu mezi helixy. Nejprve je vložena první natažená sekvence helixu, která je podrobena vkládání fragmentů za vzniku sekundární struktury. Následně dojde k připojení další sekvence helixu a proces se opakuje, dokud nejsou vloženy veškeré helixy. (B) Ukázka počátku simulace s nalezeným kontaktem. Interagující helixy jsou sbalovány v membráně spolu (převzato z Barth et al., 2009).

Energetická funkce pro membránové proteiny se skládá ze stejných částí jako pro solubilní proteiny, ale některé jsou speciálně upraveny. Například výraz pro hodnocení umístění daného typu aminokyseliny v proteinu, který v solubilním hodnocení hodnotí jen umístění na povrchu nebo uvnitř, je upraven pro kontrolu umístění v jednotlivých vrstvách membrány a její zanoření. Pokud se například daná aminokyselina vyskytuje ze 70% (zjištěné přes definované proteiny v PDB databázi) ve vnitřní hydrofobní vrstvě (viz Obr. 12) a v modelu se tam vyskytuje také, přispěje kladně do celkového hodnocení na základě dané pravděpodobnosti (Yarov-Yarovoy et al., 2006).

V membráně jsou i atypické sekundární struktury a samostatnou kapitolu tvoří zalomení helixu kvůli přítomnému prolinu (Yohannan et al., 2004). Studie ukázaly, že se narušení helixu projeví přes 4 aminokyseliny směrem k N konci řetězce (Cordes et al., 2002). Tyto ohyby jsou řešeny mimo standardní algoritmus vkládání fragmentů, kdy je využita statistická informace z ohybů určených sekvencí a jejich okolí (viz Obr. 15).



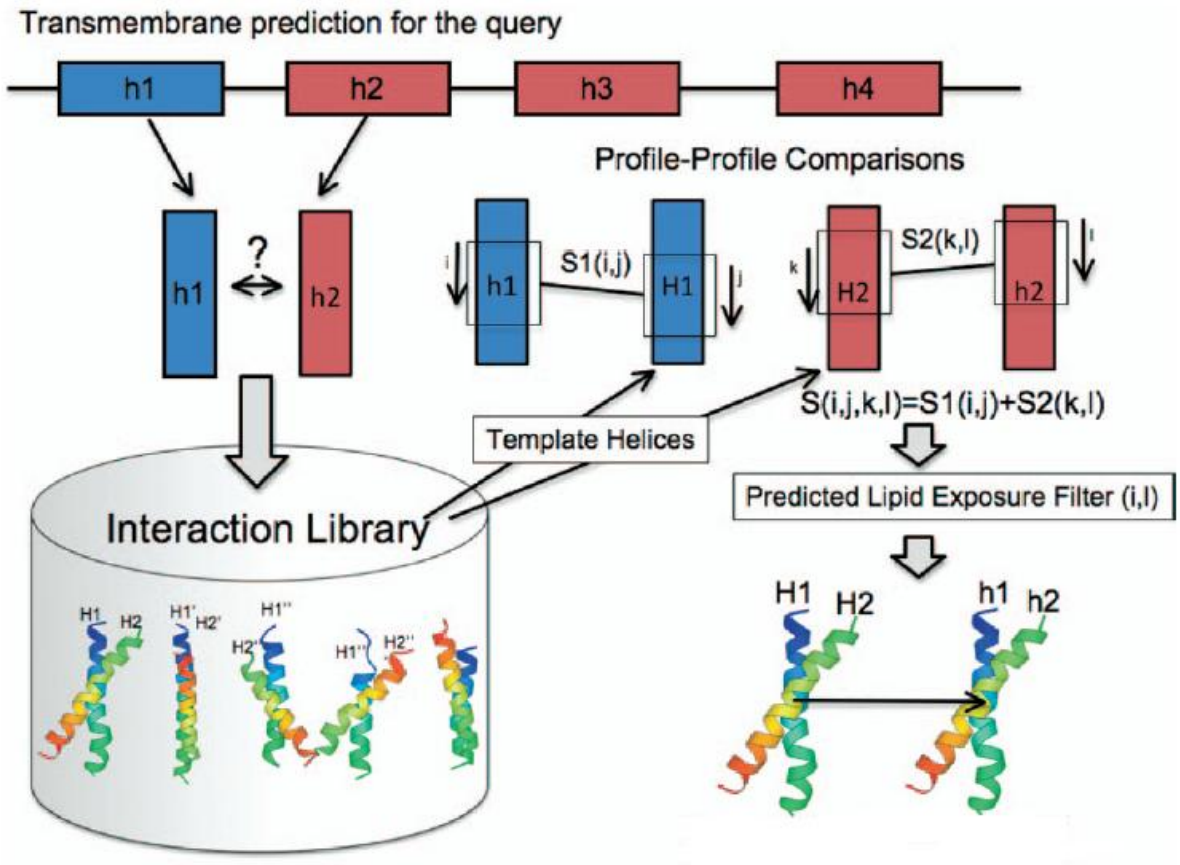
Obrázek 15: Příklad získávání informací z helixového ohybu (převzato z Barth et al., 2007)

Jelikož jsou membránové proteiny často oligomery, je možné jednotlivé predikované monomery dodatečně sloučit pomocí protokolu protein–protein docking, ale zatím samotná predikce struktury monomeru oligomerizací ovlivněná není (Barth et al., 2007).

6.7 Predikce kontaktů mezi helixy v membráně

Pro tento účel je vytvořena databáze interagujících párů helixů z experimentálně zjištěných helikálních membránových proteinů, jejichž TM helixy jsou definovány pomocí MPtopo databáze (Jayasinghe et al., 2001). Pokud mezi dvěma helixy je 5 a více C_α atomů z obou helixů v oblasti o průměru 8 Å, je tento pár zařazen do databáze. V roce 2009 k tomu bylo použito 79 proteinů s menší než 90% sekvenční identitou pro tvorbu 621 párů (Barth et al., 2009). Pomocí programu PSI-BLAST (Altschul et al., 1997) je pro každý helix v každém páru vytvořen profil jeho sekvence, který zobrazuje míru konzervovanosti jednotlivých aminokyselin (Gribskov et al., 1987).

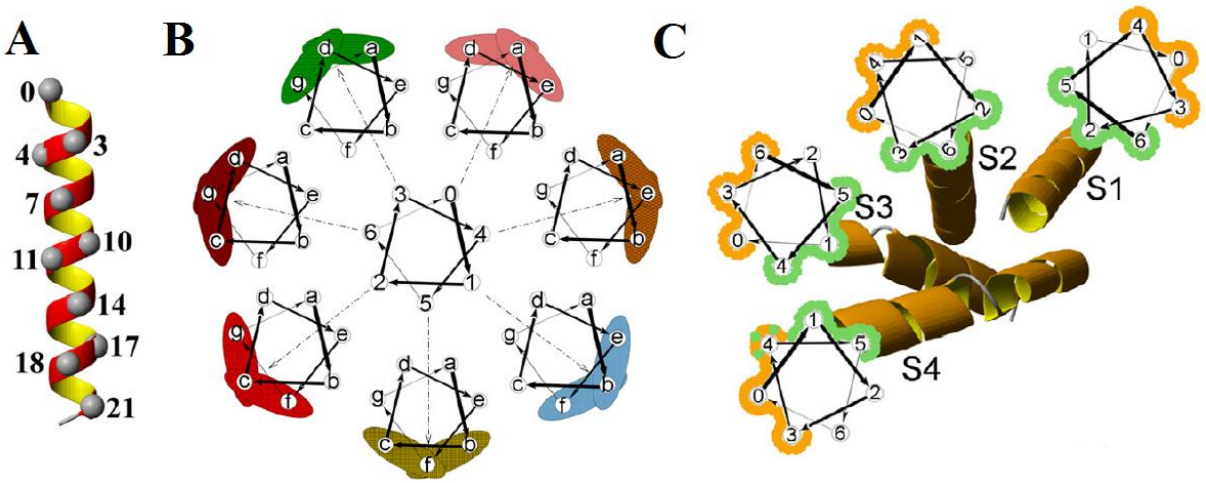
U predikované sekvence je pomocí programu OCTOPUS (Viklund a Elofsson, 2008) odhadnuta membránová topologie predikující TM helixy. Následně jsou pro každý možný helixový pár z predikované sekvence dohledány profilově odpovídající interagující páry z databáze. Dohledání je provedeno pomocí profilového porovnání přes každých 14 aminokyselin na jednotlivých helixech, a pokud dojde ke shodě, je nejbližší interagující pár aminokyselin od daného úseku definován jako kontakt (viz Obr. 16). Přes profilové porovnání jsou kontakty ohodnoceny a nejlepší vybrány pro použití na začátku predikce proteinu (Barth et al., 2009) (viz kapitola 7.6).



Obrázek 16: Hledání kontaktů (převzato z SI Barth et al., 2009)

Predikované kontakty jsou navíc verifikovány pomocí metody pro predikci vzájemné orientace TM helixů v helikálním membránovém proteinu (Adamian a Liang, 2006). Tato metoda využívá poznatky z analýz, že aminokyseliny helixů vystavené do lipidického prostředí jsou méně konzervované než ty, které směřují do proteinu a nejlépe konzervované aminokyseliny jsou ty, co interagují s nějakým kofaktorem (heme, retinal…) nebo ligandem (Adamian a Liang, 2006). Dále je použita náchylnost jednotlivých typů aminokyselin k orientaci do membrány (lipofilicita) (Adamian et al., 2005).

Vzájemná orientace TM helixů může být určena jako vyhledání plochy helixu nejvíce vystavené do membrány. Samotná definice plochy je založena na standardním modelu helixu (Pauling a Corey, 1951), kdy každá sedmá aminokyselina leží zhruba pod sebou a k doplnění je přidána navíc každá 3. a 4. (viz Obr. 17(A)). Takto definovaná plocha je navíc rozdělena na interakční části po sedmi, kdy každá sedmice je reprezentována písmeny a, b, c, d, e, f a g. Obrázek 17(B) ukazuje, jak každá sedmice aminokyselin helixu má definováno 7 interakčních částí (motif sedmi), které mohou být v kontaktu s lipidy nebo dalším helixem. Každá plocha je hodnocena pomocí funkce LIPS (LIPid-facing Surface), která určuje míru lipofilicity. Na obrázku 17(C) je zobrazena predikce 4 TM helixů.



Obrázek 17: Zjišťování lipofilicity. (A) Definice plochy helixu. (B) Každých 7 aminokyselin helixu má definováno 7 různých interakčních ploch. (C) Ukázka predikce pomocí funkce LIPS, kde oranžová plocha je predikována do membrány a zelená do proteinu (převzato z Adamian a Liang, 2006).

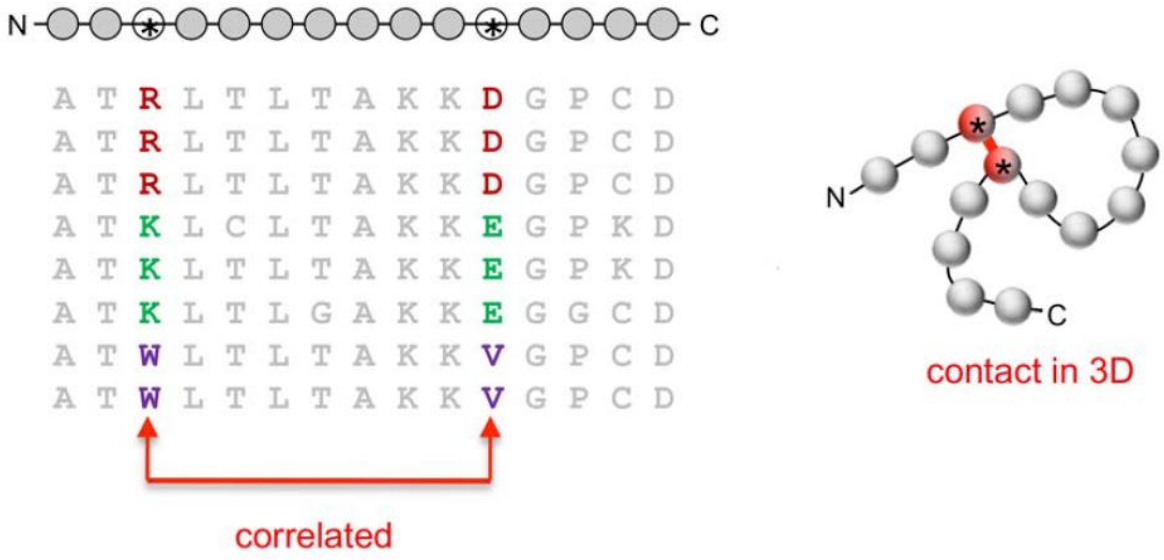
Rosetta je jeden z nejkomplexnějších nástrojů pro modelování molekulárních struktur a z několika desítek protokolů zde bylo popsáno pouze několik. Volně dostupný zdrojový kód navíc umožňuje všem možnost zavádět nové přístupy a myšlenky, a proto se Rosetta rychle rozvíjí, což dokládá i téměř 400 publikací zobrazených na oficiálních stránkách[4].

[4] https://www.rosettacommons.org/about/publications

7 EVfold

7.1 Úvod

EVfold (Evolutionary fold) je metoda založená na pozorování korelujících mutací v proteinových rodinách (Altschuh et al., 1988). Pokud se dvě aminokyseliny nacházejí v 3D prostoru přirozené konformace proteinu blízko sebe, mohou se společně evolučně měnit (Gobel et al., 1994). Toto zjištění vedlo k úvahám o použití předpovězených kontaktů aminokyselin k predikci 3D modelu (Fariselli et al., 2001; Shindyalov et al., 1994) (viz Obr. 18). Tato metoda vytvořená v roce 2011 byla schopná vytvořit testovací modely s vysokou vůči přirozené konformaci (podobností 2,7–4,8 Å C_α-RMSD) u dvou třetin testovacích proteinů (Marks et al., 2011). O rok později byla metoda upravena pro helikální membránové proteiny, kdy v testovacím souboru TM proteinů dosáhla značného úspěchu hlavně ve funkčně důležitých oblastech (Hopf et al., 2012).



Obrázek 18: Ukázka korelujícího páru aminokyselin v jedné rodině proteinů. Informace je pak použita v predikci jako pevný kontakt během simulace sbalování proteinu (převzato z Marks et al., 2011)

7.2 *Ab initio* predikce

Metoda začíná analýzou všech potenciálních párů pozic v predikované sekvenci. Pro tento účel jsou dohledány homologní sekvence pomocí databáze PFAM (Finn et al., 2010), která obsahuje množství proteinových rodin a domén. V ohodnocení korelací je nutno počítat s možností, že aminokyselina A interaguje s aminokyselinou B a ta zase s aminokyselinou C, a tím by mohlo dojít k nesprávnému párování, protože aminokyseliny A a C spolu mutačně korelují, ale nemusí spolu interagovat v 3D prostoru. Kvůli tomuto problému není analýza zaměřena striktně na dvě pozice v sekvenci, ale probíhá pomocí globálního modelu, který se snaží najít soubor párů pozic, které

nejlépe vystihují všechny pozorované korelace v homologních sekvencích. Jednotlivé páry v tomto dohledaném souboru jsou navíc ohodnoceny podle předpokládané důležitosti, která je odvozená z korelace, a nejlépe hodnocené jsou vybrány dále do predikce. Tento krok vyhledání a ohodnocení párů pozic je esenciální pro celý protokol a jeho přesnost v podstatě rozhoduje o kvalitě modelu (Marks et al., 2011).

Ve výběru jsou dále preferovány ty dvojice aminokyselin, které jsou od sebe sekvenčně vzdálenější, a také jsou upřednostňovány cysteinové můstky. Soubor vybraných párů je následovně převeden na soubor kontaktů v prostoru mezi páry aminokyselin odpovídajících pozic na predikované sekvenci. K určení vzdálenosti aminokyselin se používá C_α , C_β a odvozený střed od pozice postranního řetězce (Marks et al., 2011).

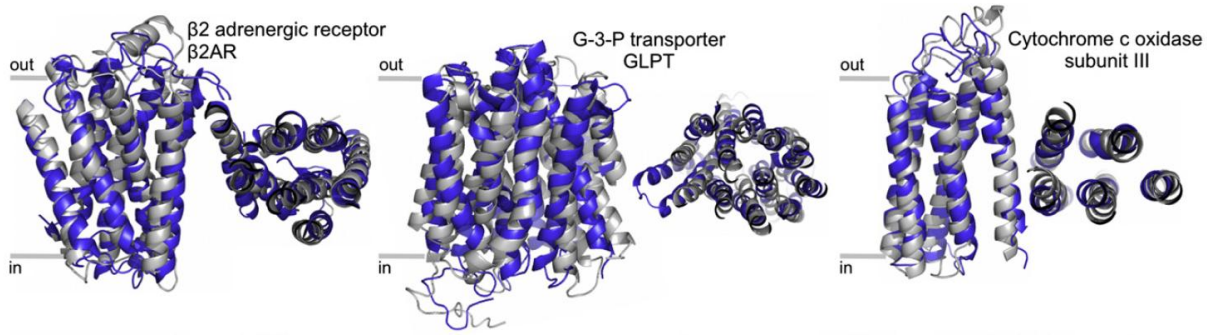
Trojrozměrný model je predikován pomocí programu CNS, který byl vytvořen pro determinaci struktury využívající data z X-ray difrakce a NMR (Brunger, 2007; Brunger et al., 1998). Tento program začíná na plně nataženém řetězci a využívá standardní „simulated annealing“ (kapitola 6.4). V rámci něho uplatňuje algoritmy zaměřené na „Distance geometry problem“. Jedná se o predikci souřadnic bodů v trojrozměrném prostoru (v našem případě jednotlivých atomů v proteinu) na základě definovaných vzdáleností mezi určitým množstvím dvojic těchto bodů (definované kontakty mezi aminokyselinami) (Havel et al., 1983).

Po vytvoření dostatečného počtu modelů jsou následně odstraněny ty, které obsahují sekvenční uzel, který dokáže nalézt veřejný server KNOT (Kolesov et al., 2007). Vytvořené modely jsou poté hodnoceny podle kvality sekundárních struktur a souhlasem s jejich predikovanými úseky (Marks et al., 2011).

7.3 EVfold_membrane

Původní algoritmus EVfold byl upraven pro predikci TM helikálních proteinů (Hopf et al., 2012). Pro nalezení homologních sekvencí byl v iniciačních testech použit program HHblits (Remmert et al., 2012) a k vytvoření 3D modelu opět použit program CNS ve verzi 1.2 (Brunger, 2007).

Počáteční výsledky testů této metody byly značně přesné (viz. Obr 19) a bylo pozorováno, že funkčně důležité úseky v proteinu jsou predikovány s větší úspěšností. Navíc úseky, kde se vyskytují nejlépe hodnocené dohledané páry pozic, jsou místa, kde dochází k vázání substrátu nebo k oligomerizaci, což může být využito při determinaci látek, které by se na tyto úseky mohly vázat, např. léky (Hopf et al., 2012).



Obrázek 19: Ukázka predikce pomocí metody EVfold_membrane, kde modrá struktura je predikovaný model a šedivá je experimentální zjištěná konformace (převzato z Hopf et al., 2012).

7.4 EVfold_bb

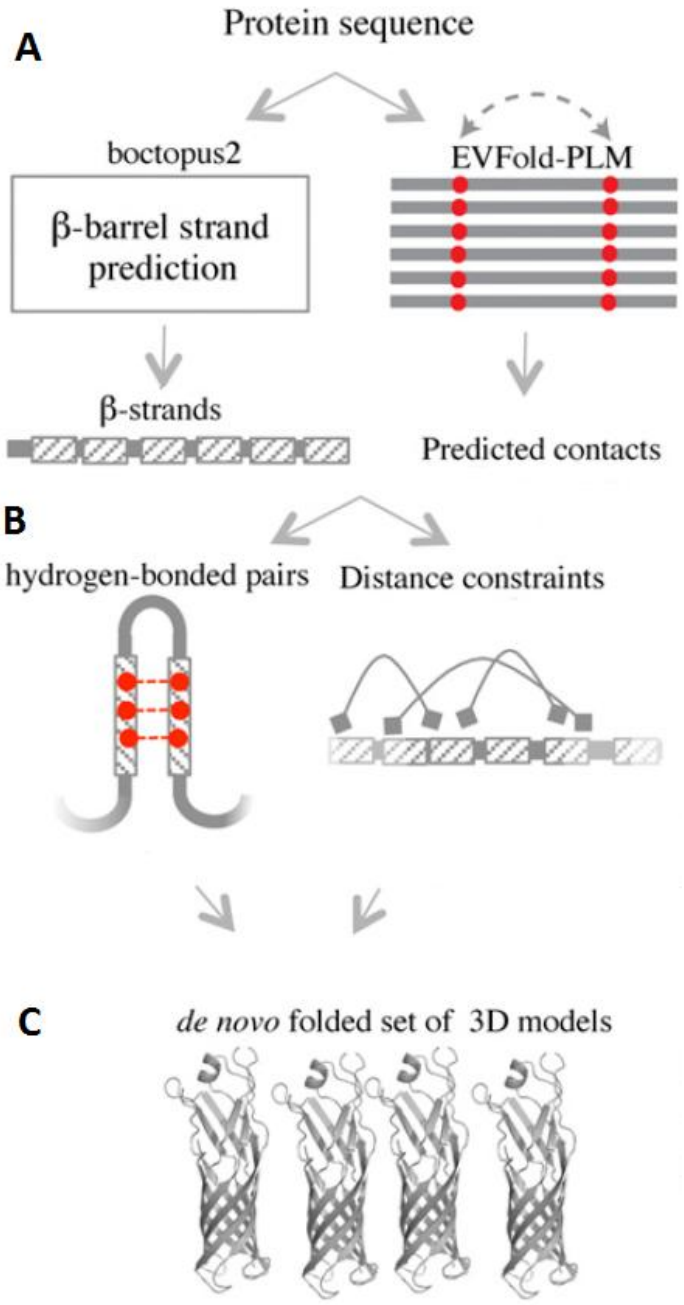
V roce 2015 byla využita metodika korelujících aminokyselin k predikci beta barelů (Hayat et al., 2015). Topologie predikované struktury je učena pomocí programu BOCTOPUS, který se zaměřuje právě na predikci TM beta barelů (Hayat a Elofsson, 2012a). V případě EVfold_bb byla použita verze 2, která každé aminokyselině určí, zda je ve smyčce na vnitřní straně membrány, vnější straně membrány nebo součástí beta řetězce v membráně. U aminokyselin v membráně navíc určí, zda je její postranní řetězec vystaven do membrány nebo je obrácen do barelu (Hayat et al., 2015) (viz Obr. 20(A)).

Korelující páry aminokyselin jsou dohledávány stejně jako ve standardním protokolu. Rozdíl je v určování vzdáleností mezi páry pro membránovou beta barelovou část a zbytkem proteinu. Vzdálenosti dvou aminokyselin v membránové části jsou určeny podle vzdálenosti dvou sousedních antiparalelních beta řetězců, kdy torzní úhly jsou nastaveny na ϕ = 135,0 a ψ = -139,0 (výchozí úhly pro antiparalelní beta list) (Hayat et al., 2015) a délky standardního vodíkového můstku (1 Å) (viz Obr. 20(B)). Vzdálenosti párů mimo membránu jsou určovány jako v kapitole 8.2.

Výsledné vzdálenosti mezi aminokyselinami jsou opět poskytnuty programu CNS (Brunger, 2007). Obrázek 20 zobrazuje kompletní postup pro tvorbu 3D modelu beta barelu (viz Obr. 20(C)).

Největší síla programu EVfold je získávání informací z pouhých sekvencí a není vázán na zjištěné struktury. Bylo by velmi zajímavé využít jeho volně dostupnou funkčnost dohledávání kontaktů a použít jí například v Rosettě. Program EVfold je možné volně vyzkoušet na jeho oficiálních stránkách [5].

[5] http://evfold.org/evfold-web/evfold.do



Obrázek 20: Diagram predikce 3D modelu beta barelu pomocí programu EVfold_bb (upraveno z Hayat et al., 2015)

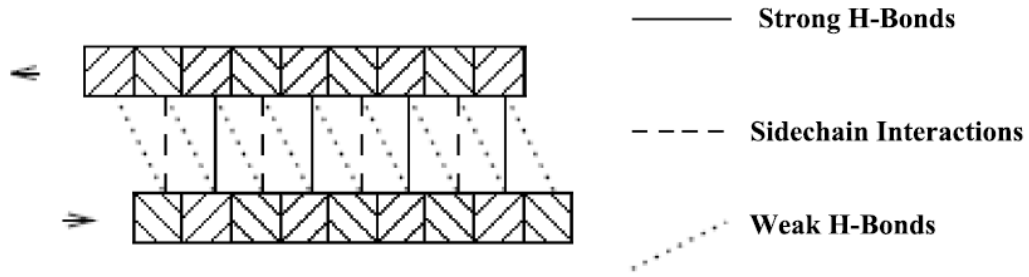
8 3D-SPOT

8.1 Úvod

3D-SPOT (3D-Structure Predictor of Transmembrane beta-barrels) je *ab initio* metoda pro predikci 3D struktury TM domény beta-barelových proteinů. Iniciační testy dokázaly vytvořit modely TM domény s RMSD peptidového řetězce standardně okolo 3,9 Å proti nativní konformaci, nicméně algoritmus má problém tvořit nesymetrické modely (viz Obr. 23), jako například PapC protein (v PDB pod kódem 2VQI). Dále je schopen predikovat, leč nepřesně, domény, které se skládají z více peptidových řetězců (Naveed et al., 2012).

8.2 *Ab initio* predikce

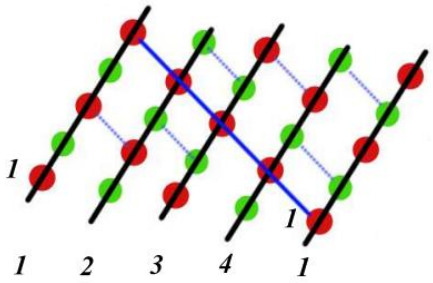
Predikce začíná zmapováním všech možných kombinací pozic, které mohou nastat mezi dvěma sousedními postranními beta řetězci. Tento proces vždy vezme jeden beta řetězec a antiparalelně k němu přiloží sekvenčně následující. Tento pár je pak vůči sobě posouván tak, aby mohlo dojít k vyhodnocení silné vodíkové vazby (Schulz, 2000), slabé vodíkové vazby (Ho a Curmi, 2002) a interakce mezi postranními řetězci (Jackups a Liang, 2005) (viz Obr. 21). K vyhodnocení těchto interakcí je použita odvozená energetická funkce, která byla vytvořena pro predikci slabých interakcí, oligomerizace a protein-proteinové interakce u beta barelů na základě experimentálně zjištěných struktur (Naveed et al., 2009). Pro každý posun je vyhodnocována i vzniklá mimo-membránová smyčka mezi beta řetězci, kdy energetická výhodnost směřuje k její menší délce (Wang et al., 2005). Takto může každý pár beta řetězců zaujmout 2L-1 poloh vůči sobě, kde L je délka řetězce (za předpokladu, že jsou délky řetězců stejné). Ty s nejmenší energií jsou následně vybrány dále pro predikci (Naveed et al., 2012).



Obrázek 21: Zobrazení dvou antiparalelních beta řetězců a jejich interakcí (upraveno z SI Naveed et al., 2012)

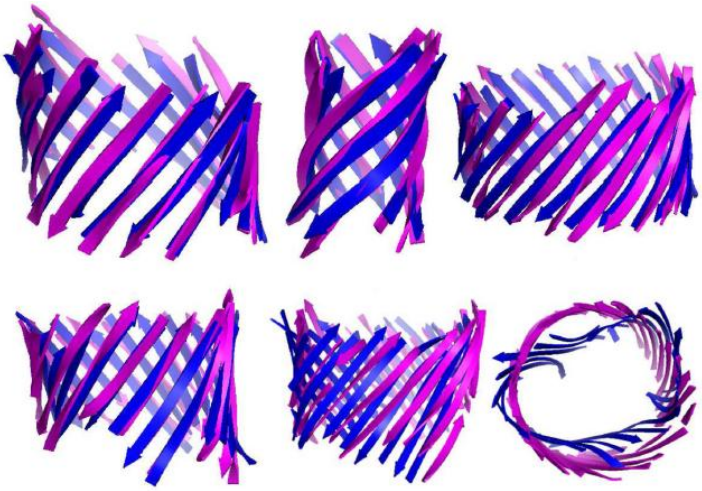
Membránová doména beta barelu většinou nabírá tvar pravidelného válce (Naveed et al., 2012), proto je tak tvořen i predikovaný model. Model je definován jako symetrický pletenec beta řetězců, kdy každý řetězec je reprezentován parametrickou křivkou a sekvencí C_α atomů. Každá pozice C_α atomu na této křivce je definována pomocí silných vodíkových můstků mezi řetězci. Průměr válce a náklon jednotlivých beta řetězců je vypočítán z počtu řetězců a střižného čísla (Shear Number), které

je vypočteno jako posun pozice aminokyseliny během celého cyklu (Liu, 1998; McLachlan, 1979) (viz Obr. 22). Toto číslo je získáno z predikovaných poloh jednotlivých řetězců s nejnižší energií a je použito pro následnou tvorbu modelu (Naveed et al., 2012).



Obrázek 22: Střižné číslo se vypočítá pomocí natažení beta barelu, kdy na začátku a na konci je ten samý řetězec s vyznačenými aminokyselinami červeně a zeleně. Ukázkový beta barel má 4 řetězce (1, 2, 3, 4) a střižné číslo 4, protože došlo k posunutí první aminokyseliny o 4 pozice (převzato z SI Naveed et al., 2012).

Šířka válce, křivky reprezentující beta řetězce a jejich náklon jsou dostatečná informace k vytvoření modelu TM domény beta barelu. Křivky jsou ale reprezentovány jen sekvencí C_α, takže je dále nutné vytvořit peptidovou kostru, a k tomu je použit program BBQ (Backbone Building from Quadrilaterals) (Gront et al., 2007). Postranní řetězce aminokyselin jsou následně dostavěny pomocí AutoPSF[6], což je rozšíření grafického programu VMD (Humphrey et al., 1996). S tím je spojeno, že každý beta řetězec může začínat s aminokyselinou, která má postranní řetězec orientovaný buď do barelu, nebo ven. V tomto kroku opět rozhoduje energetická výhodnost. Nakonec je celý model vystaven energetické minimalizaci pomocí molekulární dynamiky programem NAMD (Phillips et al., 2005).



Obrázek 23: Ukázka predikcí pomocí metody 3D-SPOT. Modré řetězce představují experimentální data, růžové predikci. Predikce zobrazená shora ukazuje problém metody při tvorbě oválných barelů u proteinu PapC (převzato z SI Naveed et al., 2012)

[6] http://www.ks.uiuc.edu/Research/vmd/plugins/autopsf/

3D-SPOT nabízí zajímavý přístup pro predikci membránové domény beta barelů. Ačkoli je jeho predikce relativně přesná, tak aktuální neschopnost tvorby nesymetrických modelů a úseků mimo membránu považuji za velký nedostatek. Snad se tyto problémy v budoucnu podaří vyřešit.

9 Ostatní programy

Kromě výše zmíněných programů uvádím ještě zmínku ostatních programů, pro jejichž detailnější popis a zhodnocení nemám dost prostoru. Tak uvádím jen velmi stručný přehled.

9.1 BCL:Fold

Metoda BCL:Fold byla vytvořena pro usnadnění predikce velkých proteinů, která je založena na kombinování idealizovaných alfa helixů a beta řetězců (Karakas et al., 2012). Tento přístup značně zmenšuje prostorové možnosti, jelikož se s predikovanými sekundárními strukturami zachází jako se samostatnými objekty, které mohou být jen lehce upravovány (například délka). V důsledku to také znamená, že se na predikovaném místě bude vždy nějaká sekundární struktura v nějaké podobě vyskytovat. Rosetta oproti tomu používá predikovaná místa sekundárních struktur jen k hodnocení po vkládání fragmentů, takže na predikovaném místě nemusí žádná sekundární struktura ve výsledném modelu být (Fischer et al., 2015; Karakas et al., 2012; Weiner et al., 2013). Je od ní odvozený membránový protokol BCL:MP-Fold, který predikuje helikální TM proteiny (Weiner et al., 2013).

9.2 TOBMODEL

TOBMODEL je metoda pro predikci TM beta barelů. Tato metoda je založena na tvorbě několika různých 3D modelů, ze kterých se vybere ten správný. Na základě predikované topologie programem BOCTOPUS (Hayat a Elofsson, 2012a) jsou vytvořeny modely, které mají různé úhly beta řetězců. Z těchto modelů je následně vybrán nejlepší pomocí ZPRED3, který predikuje vzdálenost každé aminokyseliny od centra membrány a tato predikce je porovnávána s modely (Hayat a Elofsson, 2012b).

9.3 TMBpro

TMBpro je komplexní soubor nástrojů pro predikci TM beta barelů. Zahrnuje predikci sekundárních struktur, beta kontaktů a predikci 3D modelů. Celý algoritmus na predikci 3D modelů připomíná Rosettu od vkládání fragmentů z homologních proteinů a použití „simulated annealing„ algoritmu po ohodnocení modelů energetickou funkcí (Randall et al., 2008).

10 Závěr

Moje práce je přehledem programů pro *ab initio* predikci dvou tříd TM proteinů. První třída jsou TM helikální proteiny, které se skládají hlavně ze svazku TM alfa helixů. Existují však i helikální membránové kanály a pórotvorné toxiny, které jsou velmi důležité, ale jejich predikce není plně spolehlivá. Jedním z důvodů je, že zatím nelze žádnému programu tuto informaci uživatelsky předat. Druhá třída jsou TM beta barely, které jsou převážně tvořeny sérií antiparalelních beta řetězců, které tvoří válcovitou strukturu připomínající soudek.

Ab initio predikce je zaměřena na tvorbu modelu bez použití zjištěného homologního proteinu, který by sloužil jako šablona. Tato metoda je výpočetně velmi náročná, a proto si programy musí nějak vypomoci informací z ostatních proteinů.

Rosetta tvoří model helikálního TM proteinu pomocí krátkých úseků získaných z již zjištěných proteinových struktur. Z těchto úseků je přebírána informace o jejich trojrozměrné konformaci, a tím značně redukuje možný konformační prostor, který umožňuje sekvence aminokyselin obsadit. Toto omezení je výhodné, protože odpovídá realitě, a přitom není výpočetní výkon plýtván na nereálné konformace. EVfold, který obsahuje protokol jak pro helikální TM proteiny (EVfold_membrane), tak pro beta barely (EVfold_bb), se zaměřuje na predikci kontaktů vzdálených aminokyselin, které pak použije při tvorbě trojrozměrného modelu. Tyto kontakty jsou predikovány pomocí analýzy přes sekvence proteinů ze stejné proteinové rodiny, kdy jsou dohledávány korelující mutace. Dále jsem popsal 3D-SPOT, který predikuje TM doménu beta barelu za pomocí kombinování antiparalelních sousedních beta řetězců.

Moje práce ukázala rozmanitost přístupů i problémy, které je nutné řešit. Nicméně přes složitost a komplexnost problému a malé množství programů udělala predikce TM proteinů značný pokrok.

11 Literatura

Abagyan, R. A., a S. Batalov, 1997, Do aligned sequences share the same fold?: Journal of Molecular Biology, v. 273, p. 355-368.

Adamian, L., a J. Liang, 2006, Prediction of transmembrane helix orientation in polytopic membrane proteins: Bmc Structural Biology, v. 6, p. 17.

Adamian, L., V. Nanda, W. F. DeGrado, a J. Liang, 2005, Empirical lipid propensities of amino acid residues in multispan alpha helical membrane proteins: Proteins-Structure Function a Bioinformatics, v. 59, p. 496-509.

Almen, M. S., K. J. V. Nordstrom, R. Fredriksson, a H. B. Schioth, 2009, Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function a evolutionary origin: Bmc Biology, v. 7, p. 14.

Altschuh, D., T. Vernet, P. Berti, D. Moras, a K. Nagai, 1988, COORDINATED AMINO-ACID CHANGES IN HOMOLOGOUS PROTEIN FAMILIES: Protein Engineering, v. 2, p. 193-199.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. H. Zhang, Z. Zhang, W. Miller, a D. J. Lipman, 1997, Gapped BLAST a PSI-BLAST: a new generation of protein database search programs: Nucleic Acids Research, v. 25, p. 3389-3402.

Baker, D., a A. Sali, 2001, Protein structure prediction a structural genomics: Science, v. 294, p. 93-96.

Barth, P., J. Schonbrun, a D. Baker, 2007, Toward high-resolution prediction a design of transmembrane helical protein structures: Proceedings of the National Academy of Sciences of the United States of America, v. 104, p. 15682-15687.

Barth, P., B. Wallner, a D. Baker, 2009, Prediction of membrane protein structures with complex topologies using limited constraints: Proceedings of the National Academy of Sciences of the United States of America, v. 106, p. 1409-1414.

Bateman, A., L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. L. Sonnhammer, D. J. Studholme, C. Yeats, a S. R. Eddy, 2004, The Pfam protein families database: Nucleic Acids Research, v. 32, p. D138-D141.

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, a P. E. Bourne, 2000, The Protein Data Bank: Nucleic Acids Research, v. 28, p. 235-242.

Bill, R. M., P. J. F. Henderson, S. Iwata, E. R. S. Kunji, H. Michel, R. Neutze, S. Newstead, B. Poolman, C. G. Tate, a H. Vogel, 2011, Overcoming barriers to membrane protein structure determination: Nature Biotechnology, v. 29, p. 335-340.

Bishop, R. E., 2008, Structural biology of membrane-intrinsic beta-barrel enzymes: Sentinels of the bacterial outer membrane: Biochimica Et Biophysica Acta-Biomembranes, v. 1778, p. 1881-1896.

Bonneau, R., I. Ruczinski, J. Tsai, a D. Baker, 2002, Contact order a ab initio protein structure prediction: Protein Science, v. 11, p. 1937-1944.

Bonneau, R., C. E. M. Strauss, a D. Baker, 2001a, Improving the performance of Rosetta using multiple sequence alignment information a global measures of hydrophobic core formation: Proteins-Structure Function a Bioinformatics, v. 43, p. 1-11.

Bonneau, R., J. Tsai, I. Ruczinski, D. Chivian, C. Rohl, C. E. M. Strauss, a D. Baker, 2001b, Rosetta in CASP4: Progress in ab initio protein structure prediction: Proteins-Structure Function a Genetics, p. 119-126.

Bradley, P., a D. Baker, 2006, Improved beta-protein structure prediction by multilevel optimization of NonLocal strand pairings a local backbone conformation: Proteins-Structure Function a Bioinformatics, v. 65, p. 922-929.

Bradley, P., D. Chivian, J. Meiler, K. M. S. Misura, C. A. Rohl, W. R. Schief, W. J. Wedemeyer, O. Schueler-Furman, P. Murphy, J. Schonbrun, C. E. M. Strauss, a D. Baker, 2003, Rosetta predictions in CASP5: Successes, failures, a prospects for complete automation: Proteins-Structure Function a Genetics, v. 53, p. 457-468.

Bradley, P., L. Malmstrom, B. Qian, J. Schonbrun, D. Chivian, D. E. Kim, K. Meiler, K. M. S. Misura, a D. Baker, 2005, Free modeling with Rosetta in CASP6: Proteins-Structure Function a Bioinformatics, v. 61, p. 128-134.

Brenner, S. E., C. Chothia, a T. J. P. Hubbard, 1998, Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships: Proceedings of the National Academy of Sciences of the United States of America, v. 95, p. 6073-6078.

Brunger, A. T., 2007, Version 1.2 of the Crystallography a NMR system: Nature Protocols, v. 2, p. 2728-2733.

Brunger, A. T., P. D. Adams, G. M. Clore, W. L. DeLano, P. Gros, R. W. Grosse-Kunstleve, J. S. Jiang, J. Kuszewski, M. Nilges, N. S. Pannu, R. J. Read, L. M. Rice, T. Simonson, a G. L. Warren, 1998, Crystallography & NMR system: A new software suite for macromolecular structure determination: Acta Crystallographica Section D-Biological Crystallography, v. 54, p. 905-921.

Bryngelson, J. D., J. N. Onuchic, N. D. Socci, a P. G. Wolynes, 1995, FUNNELS, PATHWAYS, a THE ENERGY LANDSCAPE OF PROTEIN-FOLDING - A SYNTHESIS: Proteins-Structure Function a Genetics, v. 21, p. 167-195.

Carugo, O., a S. Pongor, 2001, A normalized root-mean-square distance for comparing protein three-dimensional structures: Protein Science, v. 10, p. 1470-1473.

Cassarino, T. G., L. Bordoli, a T. Schwede, 2014, Assessment of ligand binding site predictions in CASP10: Proteins-Structure Function a Bioinformatics, v. 82, p. 154-163.

Cordes, F. S., J. N. Bright, a M. S. P. Sansom, 2002, Proline-induced distortions of transmembrane helices: Journal of Molecular Biology, v. 323, p. 951-960.

Coutsias, E. A., C. Seok, a K. A. Dill, 2004, Using quaternions to calculate RMSD: Journal of Computational Chemistry, v. 25, p. 1849-1857.

Delcour, A. H., 2002, Structure a function of pore-forming beta-barrels from bacteria: Journal of Molecular Microbiology a Biotechnology, v. 4, p. 1-10.

Dunbrack, R. L., a F. E. Cohen, 1997, Bayesian statistical analysis of protein side-chain rotamer preferences: Protein Science, v. 6, p. 1661-1681.

Durham, E., B. Dorr, N. Woetzel, R. Staritzbichler, a J. Meiler, 2009, Solvent accessible surface area approximations for rapid a accurate protein structure prediction: Journal of Molecular Modeling, v. 15, p. 1093-1108.

Eilers, M., A. B. Patel, W. Liu, a S. O. Smith, 2002, Comparison of helix interactions in membrane a soluble alpha-bundle proteins: Biophysical Journal, v. 82, p. 2720-2736.

Fagerberg, L., K. Jonasson, G. von Heijne, M. Uhlen, a L. Berglund, 2010, Prediction of the human membrane proteome: Proteomics, v. 10, p. 1141-1149.

Fariselli, P., O. Olmea, A. Valencia, a R. Casadio, 2001, Prediction of contact maps with neural networks a correlated mutations: Protein Engineering, v. 14, p. 835-843.

Finn, R. D., J. Mistry, J. Tate, P. Coggill, A. Heger, J. E. Pollington, O. L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E. L. L. Sonnhammer, S. R. Eddy, a A. Bateman, 2010, The Pfam protein families database: Nucleic Acids Research, v. 38, p. D211-D222.

Fischer, A. W., N. S. Alexander, N. Woetzel, M. Karakas, B. E. Weiner, a J. Meiler, 2015, BCL::MP-fold: Membrane protein structure prediction guided by EPR restraints: Proteins-Structure Function a Bioinformatics, v. 83, p. 1947-1962.

Freeman, T. C., a W. C. Wimley, 2010, A highly accurate statistical approach for the prediction of transmembrane beta-barrels: Bioinformatics, v. 26, p. 1965-1974.

Gobel, U., C. Sander, R. Schneider, a A. Valencia, 1994, CORRELATED MUTATIONS a RESIDUE CONTACTS IN PROTEINS: Proteins-Structure Function a Genetics, v. 18, p. 309-317.

Grantcharova, V., E. J. Alm, D. Baker, a A. L. Horwich, 2001, Mechanisms of protein folding: Current Opinion in Structural Biology, v. 11, p. 70-82.

Gray, J. J., S. Moughon, C. Wang, O. Schueler-Furman, B. Kuhlman, C. A. Rohl, a D. Baker, 2003, Protein-protein docking with simultaneous optimization of rigid-body displacement a side-chain conformations: Journal of Molecular Biology, v. 331, p. 281-299.

Gribskov, M., A. D. McLachlan, a D. Eisenberg, 1987, PROFILE ANALYSIS - DETECTION OF DISTANTLY RELATED PROTEINS: Proceedings of the National Academy of Sciences of the United States of America, v. 84, p. 4355-4358.

Gront, D., S. Kmiecik, a A. Kolinski, 2007, Backbone building from quadrilaterals: A fast a accurate algorithm for protein backbone reconstruction from alpha carbon coordinates: Journal of Computational Chemistry, v. 28, p. 1593-1597.

Gunn, J., 1998, Hierarchical minimization with distance a angle constraints.

Havel, T. F., I. D. Kuntz, a G. M. Crippen, 1983, THE COMBINATORIAL DISTANCE GEOMETRY METHOD FOR THE CALCULATION OF MOLECULAR-CONFORMATION .1. A NEW APPROACH TO AN OLD PROBLEM: Journal of Theoretical Biology, v. 104, p. 359-381.

Hayat, S., a A. Elofsson, 2012a, BOCTOPUS: improved topology prediction of transmembrane beta barrel proteins: Bioinformatics, v. 28, p. 516-522.

Hayat, S., a A. Elofsson, 2012b, Ranking models of transmembrane beta-barrel proteins using Z-coordinate predictions: Bioinformatics, v. 28, p. I90-I96.

Hayat, S., C. Sander, D. S. Marks, a A. Elofsson, 2015, All-atom 3D structure prediction of transmembrane beta-barrel proteins from sequences: Proceedings of the National Academy of Sciences of the United States of America, v. 112, p. 5413-5418.

Ho, B. K., a P. M. G. Curmi, 2002, Twist a shear in beta-sheets a beta-ribbons: Journal of Molecular Biology, v. 317, p. 291-308.

Hobohm, U., M. Scharf, R. Schneider, a C. Sander, 1992, SELECTION OF REPRESENTATIVE PROTEIN DATA SETS: Protein Science, v. 1, p. 409-417.

Hopf, T. A., L. J. Colwell, R. Sheridan, B. Rost, C. Sander, a D. S. Marks, 2012, Three-Dimensional Structures of Membrane Proteins from Genomic Sequencing: Cell, v. 149, p. 1607-1621.

Humphrey, W., A. Dalke, a K. Schulten, 1996, VMD: Visual molecular dynamics: Journal of Molecular Graphics & Modelling, v. 14, p. 33-38.

Irving, J. A., J. C. Whisstock, a A. M. Lesk, 2001, Protein structural alignments a functional genomics: Proteins-Structure Function a Genetics, v. 42, p. 378-382.

Jackups, R., a J. Liang, 2005, Interstrand pairing patterns in beta-barrel membrane proteins: The positive-outside rule, aromatic rescue, a strand registration prediction: Journal of Molecular Biology, v. 354, p. 979-993.

Javadpour, M. M., M. Eilers, M. Groesbeek, a S. O. Smith, 1999, Helix packing in polytopic membrane proteins: Role of glycine in transmembrane helix association: Biophysical Journal, v. 77, p. 1609-1618.

Jayasinghe, S., K. Hristova, a S. H. White, 2001, MPtopo: A database of membrane protein topology: Protein Science, v. 10, p. 455-458.

Jones, D. T., 1999, Protein secondary structure prediction based on position-specific scoring matrices: Journal of Molecular Biology, v. 292, p. 195-202.

Kabsch, W., 1976, SOLUTION FOR BEST ROTATION TO RELATE 2 SETS OF VECTORS: Acta Crystallographica Section A, v. 32, p. 922-923.

Karakas, M., N. Woetzel, R. Staritzbichler, N. Alexander, B. E. Weiner, a J. Meiler, 2012, BCL::Fold - De Novo Prediction of Complex a Large Protein Topologies by Assembly of Secondary Structure Elements: Plos One, v. 7, p. 20.

Karplus, K., C. Barrett, M. Cline, M. Diekhans, L. Grate, a R. Hughey, 1999, Predicting protein structure using only sequence information: Proteins-Structure Function a Genetics, p. 121-125.

Karplus, K., R. Karchin, C. Barrett, S. Tu, M. Cline, M. Diekhans, L. Grate, J. Casper, a R. Hughey, 2001, What is the value added by human intervention in protein structure prediction?: Proteins-Structure Function a Genetics, p. 86-91.

Kirkpatrick, S., C. D. Gelatt, a M. P. Vecchi, 1983, OPTIMIZATION BY SIMULATED ANNEALING: Science, v. 220, p. 671-680.

Kolesov, G., P. Virnau, M. Kardar, a L. A. Mirny, 2007, Protein knot server: detection of knots in protein structures: Nucleic Acids Research, v. 35, p. W425-W428.

Kryshtafovych, A., A. Barbato, K. Fidelis, B. Monastyrskyy, T. Schwede, a A. Tramontano, 2014, Assessment of the assessment: Evaluation of the model quality estimates in CASP10: Proteins-Structure Function a Bioinformatics, v. 82, p. 112-126.

Kuhlman, B., a D. Baker, 2000, Native protein sequences are close to optimal for their structures: Proceedings of the National Academy of Sciences of the United States of America, v. 97, p. 10383-10388.

Larson, S. M., a A. R. Davidson, 2000, The identification of conserved interactions within the SH3 domain by alignment of sequences a structures: Protein Science, v. 9, p. 2170-2180.

Leman, J. K., M. B. Ulmschneider, a J. J. Gray, 2015, Computational modeling of membrane proteins: Proteins-Structure Function a Bioinformatics, v. 83, p. 1-24.

Lindahl, E., a M. S. P. Sansom, 2008, Membrane proteins: molecular dynamics simulations: Current Opinion in Structural Biology, v. 18, p. 425-431.

Liu, W. M., 1998, Shear numbers of protein beta-barrels: Definition refinements a statistics: Journal of Molecular Biology, v. 275, p. 541-545.

Marks, D. S., L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, a C. Sander, 2011, Protein 3D Structure Computed from Evolutionary Sequence Variation: Plos One, v. 6, p. 20.

McLachlan, A. D., 1979, GENE DUPLICATIONS IN THE STRUCTURAL EVOLUTION OF CHYMOTRYPSIN: Journal of Molecular Biology, v. 128, p. 49-&.

Meiler, J., M. Mueller, A. Zeidler, a e. al., 2002, JUFO: secondary structure prediction for proteins.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, a E. Teller, 1953, EQUATION OF STATE CALCULATIONS BY FAST COMPUTING MACHINES: Journal of Chemical Physics, v. 21, p. 1087-1092.

Misura, K. M. S., a D. Baker, 2005, Progress a challenges in high-resolution refinement of protein structure models: Proteins-Structure Function a Bioinformatics, v. 59, p. 15-29.

Monastyrskyy, B., D. D'Andrea, K. Fidelis, A. Tramontano, a A. Kryshtafovych, 2014a, Evaluation of residue-residue contact prediction in CASP10: Proteins-Structure Function a Bioinformatics, v. 82, p. 138-153.

Monastyrskyy, B., A. Kryshtafovych, J. Moult, A. Tramontano, a K. Fidelis, 2014b, Assessment of protein disorder region predictions in CASP10: Proteins-Structure Function a Bioinformatics, v. 82, p. 127-137.

Moult, J., K. Fidelis, A. Kryshtafovych, T. Schwede, a A. Tramontano, 2014, Critical assessment of methods of protein structure prediction (CASP) - round x: Proteins-Structure Function a Bioinformatics, v. 82, p. 1-6.

Naveed, H., R. Jackups, a J. Liang, 2009, Predicting weakly stable regions, oligomerization state, a protein-protein interfaces in transmembrane domains of outer membrane proteins: Proceedings of the National Academy of Sciences of the United States of America, v. 106, p. 12735-12740.

Naveed, H., Y. Xu, R. Jackups, a J. Liang, 2012, Predicting Three-Dimensional Structures of Transmembrane Domains of beta-Barrel Membrane Proteins: Journal of the American Chemical Society, v. 134, p. 1775-1781.

Nugent, T., D. Cozzetto, a D. T. Jones, 2014, Evaluation of predictions in the CASP10 model refinement category: Proteins-Structure Function a Bioinformatics, v. 82, p. 98-111.

Overington, J. P., B. Al-Lazikani, a A. L. Hopkins, 2006, Opinion - How many drug targets are there?: Nature Reviews Drug Discovery, v. 5, p. 993-996.

Pauling, L., a R. B. Corey, 1951, THE STRUCTURE OF SYNTHETIC POLYPEPTIDES: Proceedings of the National Academy of Sciences of the United States of America, v. 37, p. 241-250.

Phillips, J. C., R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, a K. Schulten, 2005, Scalable molecular dynamics with NAMD: Journal of Computational Chemistry, v. 26, p. 1781-1802.

Plaxco, K. W., K. T. Simons, a D. Baker, 1998, Contact order, transition state placement a the refolding rates of single domain proteins: Journal of Molecular Biology, v. 277, p. 985-994.

Randall, A., J. L. Cheng, M. Sweredoski, a P. Baldi, 2008, TMBpro: secondary structure, beta-contact a tertiary structure prediction of transmembrane beta-barrel proteins: Bioinformatics, v. 24, p. 513-520.

Reeb, J., E. Kloppmann, M. Bernhofer, a B. Rost, 2015, Evaluation of transmembrane helix predictions in 2014: Proteins-Structure Function a Bioinformatics, v. 83, p. 473-484.

Rees, D. C., L. Deantonio, a D. Eisenberg, 1989, HYDROPHOBIC ORGANIZATION OF MEMBRANE-PROTEINS: Science, v. 245, p. 510-513.

Remmert, M., A. Biegert, A. Hauser, a J. Soding, 2012, HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment: Nature Methods, v. 9, p. 173-175.

Rohl, C. A., C. E. M. Strauss, K. M. S. Misura, a D. Baker, 2004, Protein structure prediction using rosetta: Numerical Computer Methods, Pt D, v. 383, p. 66-+.

Sander, C., a R. Schneider, 1991, DATABASE OF HOMOLOGY-DERIVED PROTEIN STRUCTURES a THE STRUCTURAL MEANING OF SEQUENCE ALIGNMENT: Proteins-Structure Function a Genetics, v. 9, p. 56-68.

Schleiff, E., a J. Soll, 2005, Membrane protein insertion: mixing eukaryotic a prokaryotic concepts: Embo Reports, v. 6, p. 1023-1027.

Schulz, G. E., 2000, beta-Barrel membrane proteins: Current Opinion in Structural Biology, v. 10, p. 443-447.

Shaw, D. E., P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. B. Shan, a W. Wriggers, 2010, Atomic-Level Characterization of the Structural Dynamics of Proteins: Science, v. 330, p. 341-346.

Shindyalov, I. N., N. A. Kolchanov, a C. Sander, 1994, CAN 3-DIMENSIONAL CONTACTS IN PROTEIN STRUCTURES BE PREDICTED BY ANALYSIS OF CORRELATED MUTATIONS: Protein Engineering, v. 7, p. 349-358.

Shortle, D., K. T. Simons, a D. Baker, 1998, Clustering of low-energy conformations near the native structures of small proteins: Proceedings of the National Academy of Sciences of the United States of America, v. 95, p. 11158-11162.

Simons, K. T., R. Bonneau, I. Ruczinski, a D. Baker, 1999a, Ab initio protein structure prediction of CASP III targets using ROSETTA: Proteins-Structure Function a Bioinformatics, p. 171-176.

Simons, K. T., C. Kooperberg, E. Huang, a D. Baker, 1997, Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing a Bayesian scoring functions: Journal of Molecular Biology, v. 268, p. 209-225.

Simons, K. T., I. Ruczinski, C. Kooperberg, B. A. Fox, C. Bystroff, a D. Baker, 1999b, Improved recognition of native-like protein structures using a combination of sequence-dependent a sequence-independent features of proteins: Proteins-Structure Function a Genetics, v. 34, p. 82-95.

Song, L. Z., M. R. Hobaugh, C. Shustak, S. Cheley, H. Bayley, a J. E. Gouaux, 1996, Structure of staphylococcal alpha-hemolysin, a heptameric transmembrane pore: Science, v. 274, p. 1859-1866.

Taylor, T. J., H. J. Bai, C. H. Tai, a B. Lee, 2014, Assessment of CASP10 contact-assisted predictions: Proteins-Structure Function a Bioinformatics, v. 82, p. 84-97.

Tsai, J., R. Bonneau, A. V. Morozov, B. Kuhlman, C. A. Rohl, a D. Baker, 2003, An improved protein decoy set for testing energy functions for protein structure prediction: Proteins-Structure Function a Genetics, v. 53, p. 76-87.

Tusnady, G. E., Z. Dosztanyi, a I. Simon, 2004, Transmembrane proteins in the Protein Data Bank: identification a classification: Bioinformatics, v. 20, p. 2964-2972.

Ulmschneider, M. B., M. S. P. Sansom, a A. Di Nola, 2005, Properties of integral membrane protein structures: Derivation of an implicit membrane potential: Proteins-Structure Function a Bioinformatics, v. 59, p. 252-265.

Viklund, H., a A. Elofsson, 2008, OCTOPUS: improving topology prediction by two-track ANN-based preference scores a an extended topological grammar: Bioinformatics, v. 24, p. 1662-1668.

Waldispuehl, J., C. W. O'Donnell, S. Devadas, P. Clote, a B. Berger, 2008, Modeling ensembles of transmembrane beta-barrel proteins: Proteins-Structure Function a Bioinformatics, v. 71, p. 1097-1112.

Wang, L. P., E. V. Rivera, M. G. Benavides-Garcia, a B. T. Nall, 2005, Loop entropy a cytochrome c stability: Journal of Molecular Biology, v. 353, p. 719-729.

Weiner, B. E., N. Woetzel, M. Karakas, N. Alexander, a J. Meiler, 2013, BCL::MP-Fold: Folding Membrane Proteins through Assembly of Transmembrane Helices: Structure, v. 21, p. 1107-1117.

White, S. H., a W. C. Wimley, 1999, Membrane protein folding a stability: Physical principles: Annual Review of Biophysics a Biomolecular Structure, v. 28, p. 319-365.

Xu, D., a R. Nussinov, 1998, Favorable domain size in proteins: Folding & Design, v. 3, p. 11-17.

Yarov-Yarovoy, V., J. Schonbrun, a D. Baker, 2006, Multipass membrane protein structure prediction using Rosetta: Proteins-Structure Function a Bioinformatics, v. 62, p. 1010-1025.

Yohannan, S., S. Faham, D. Yang, J. P. Whitelegge, a J. U. Bowie, 2004, The evolution of transmembrane helix kinks a the structural diversity of G protein-coupled receptors: Proceedings of the National Academy of Sciences of the United States of America, v. 101, p. 959-963.